ST 2/3

Health Statistics

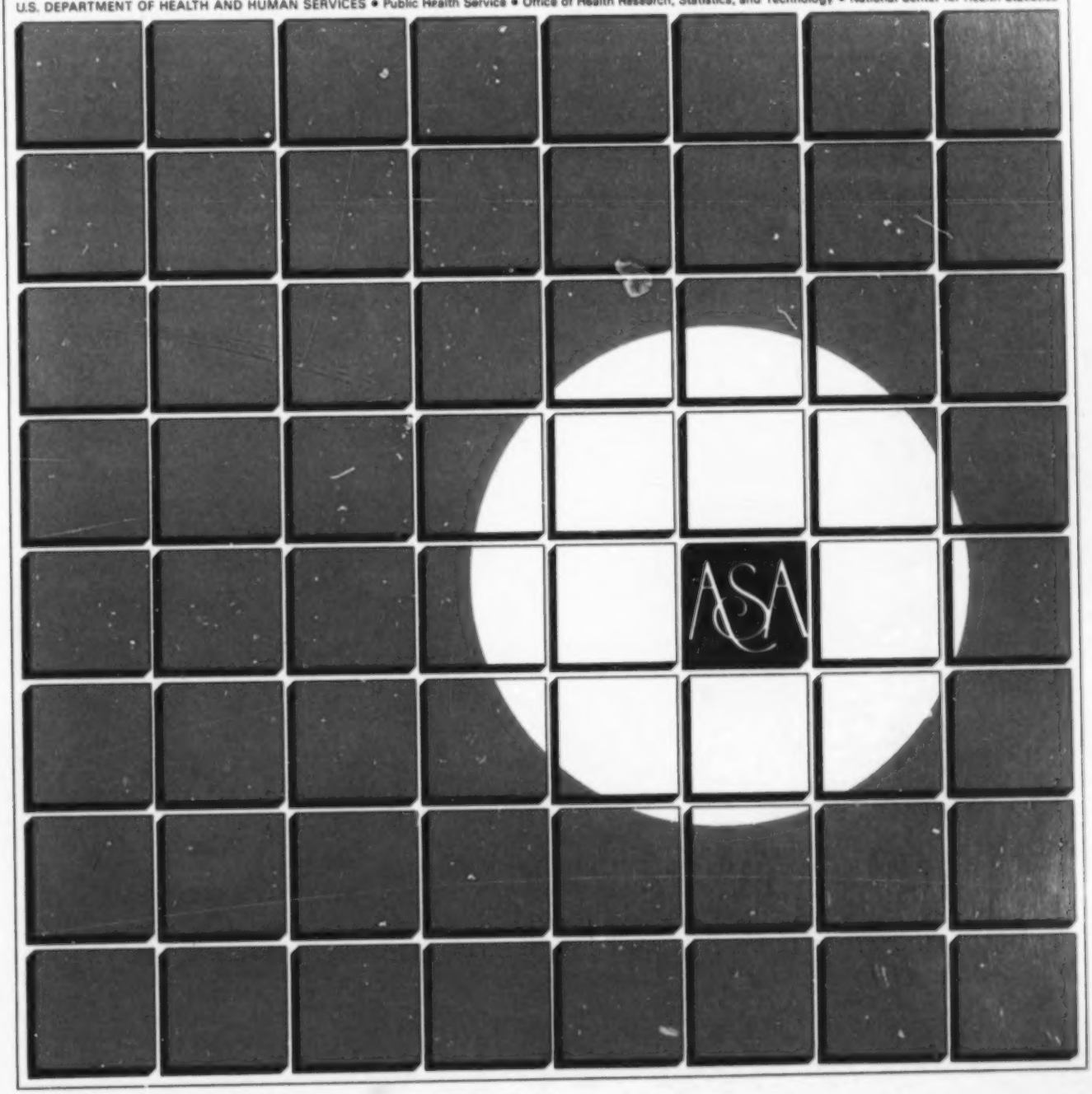
Contributed Papers Session
American Statistical Association
August 1982

NEWARK PUBLIC LIBRARY

SEP 17 1982

U. S. GOVT. PUBS.
DEPOSITORY 376

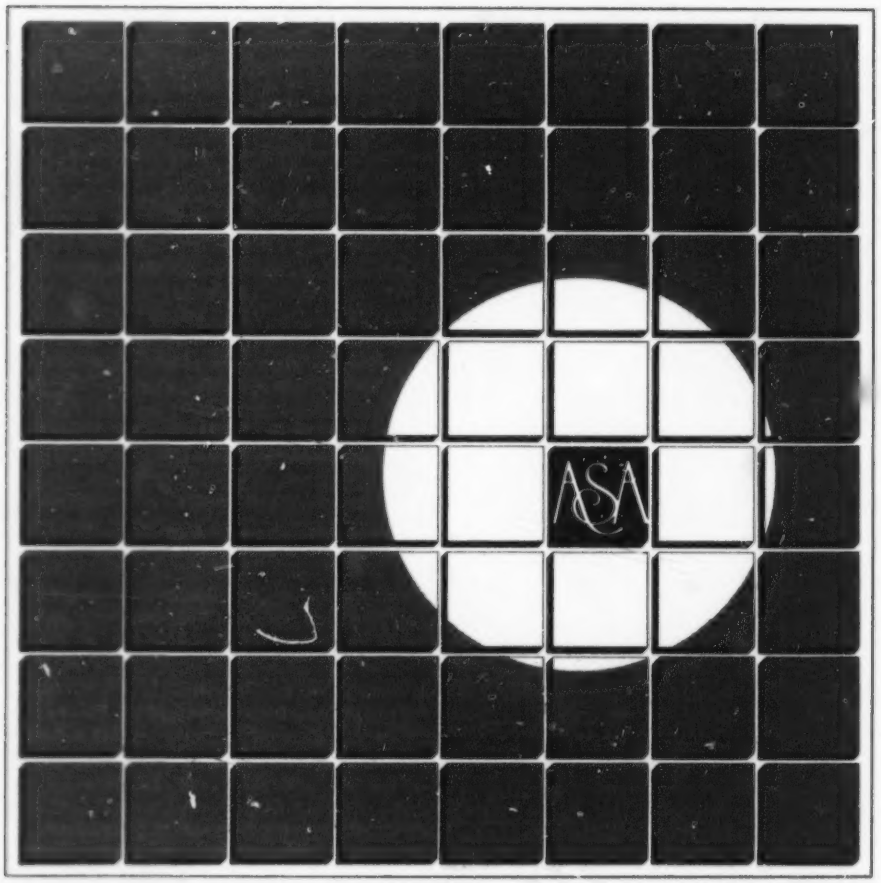
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES • Public Health Service • Office of Health Research, Statistics, and Technology • National Center for Health Statistics



HE 20.6202

Health Statistics

Contributed Papers Session
American Statistical Association
August 1982



U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Office of Health Research, Statistics, and Technology
National Center for Health Statistics

Hyattsville, Maryland
August 1982

READEX FICHE

CONTENTS

Chair--Mary Grace Kovar

Preface.....v

A Model for Designing NCHS's Population Based Surveys--James T. Massey and Monroe G. Sirken.....1

Stratifying Primary Sampling Units with the SAS Cluster Procedure --Joe Fred Gonzalez and Andrew A. White.....7

NCHS Research on the Telephone Interview: Some Observations --Owen T. Thornberry, Jr., and Gail S. Poe.....12

Indirect Standardization of Neonatal Mortality for Birth Weight --Joel C. Kleinman.....25

Monthly Patterns of Marital Relations and Waiting Times to Conception--Gerry E. Hendershot.....36

Discussant--Leslie Kish

A MODEL FOR DESIGNING NCHS'S POPULATION BASED SURVEYS

James T. Massey and Monroe G. Sirken
National Center for Health Statistics

1. INTRODUCTION

The National Health Interview Survey (NHIS) is a continuous weekly survey whose major objective is to produce national statistics of morbidity, disability, and health services utilization. Since the NHIS was first implemented in 1957, the sample design has been reevaluated and updated following each decennial census of the population. The redesigns following the 1960 and 1970 censuses were for the most part updates of the sampling frame to reflect changes in the population characteristics that had occurred between censuses.

The redesign of the NHIS following the 1980 census is scheduled to be implemented in 1985. During the past several years a considerable effort has been made by NCHS with the assistance of the Bureau of the Census to take an in-depth look at the current and future objectives and data requirements of the NHIS and to thoroughly reevaluate the survey design of the NHIS. The redesign period is also being used to incorporate improved survey methodologies into the NHIS design. At the same time new survey methodologies are being considered for the NHIS, the survey designs for the Center's other population based surveys are being evaluated. Traditionally, the NHIS, the National Medical Care Expenditure Survey (NMCUES), the National Survey of Family Growth (NSFG), and the National Health and Nutrition Examination Survey (NHANES) have been designed and implemented independently. For this redesign, the possibility of redesigning the NHIS so that it could serve as a sampling frame for the other surveys is being explored. In undertaking such a complex task as selecting a single design for several surveys, a model was used to identify and evaluate a number of design features. This paper describes the redesign process and reports on the progress to date.

2. OBJECTIVES OF THE REDESIGN

In redesigning an existing survey it is imperative that the objectives of the survey be clearly specified. This is sometimes a more difficult task than stating a set of objectives for a new survey. Plewes [1] states that "possibly the only task more challenging to the survey statistician than initially designing a large-scale socioeconomic survey is that of redesigning one." Since the redesign involves an ongoing program, many resources have already been invested in the current design, and a shift in objectives often puts the old objectives in conflict with new objectives. The background and objectives for the NHIS redesign were presented by Burnham and Massey [2] at the 1980 ASA meetings. The foremost objective for the NHIS continues to be the production of national descriptive health statistics by socioeconomic and sociodemographic variables. The additional objectives for the redesign are:

1) to improve the precision of the NHIS estimates for special small domains and for estimates of change,

2) to improve the efficiency and analytical utility of the Center's combined population survey program, and

3) to reduce the cost of data collection for the Center's population based surveys by introducing new survey methodologies.

Although the NHIS has always been a multipurpose survey which collected information on a variety of health topics, the expanded objectives for the 1980's presented NCHS with an almost overwhelming task of selecting a single sample design which would be able to satisfy all or most of these major objectives. In order to translate these objectives into an operational design strategy, a model approach was used to help evaluate different design alternatives.

3. DESIGN MODEL

The model used for the redesign was first developed by Sirken and Roysten [3] to study the design effects in retrospective mortality surveys. We have further refined the model and applied it to the redesign of the Center's population based surveys.

Conceptually, the use of a model to define the design process is straight forward. The first step is to identify the major design features related to the survey design. A design feature is defined as a destinctive part of the survey design that is manipulatable. Examples of design features are the questionnaire, the sample design parameters such as the stratification variables and number of primary sampling units, the sampling frame, and the mode of data collection.

Once the design features have been enumerated the possible design options associated with each feature are identified. One of the possible options is the null option which implies the elimination of the design feature. Examples of the design options for the design feature, mode of data collection, are face-to-face interviews, telephone interviews, and self-administered questionnaires. Rather than enumerate all possible options associated with each design feature, we shall restrict our model to the feasible design options. In the NHIS a mail questionnaire would never be used because of the complexity of the NHIS questionnaire and therefore represents an infeasible design option in the example given above.

After the design options have been defined for each of the design features, design option sets

are formed. A design option set is composed of a single design option from each of the design features. A design option set represents a possible alternative survey design. All possible design option sets are then enumerated. The number of design option sets is the product of the number of design options for each feature.

Following enumeration, the design option sets are systematically evaluated in two phases. During the first phase the design options within a design option set are examined for compatibility with each other. If any two design options are incompatible the design option set is deleted. The second phase of the evaluation involves the application of a set of critera to identify the optimum design option set. The optimum design option set represents the survey design to be implemented. A single criterion might be the minimum sampling variance or mean squared error for a fixed cost associated with each of the design option sets. The variance or mean squared error might represent an average value over a number of variables of interest in the survey. Another criteria might be an evaluation of how well each of the design option sets satisfies the major objectives of the survey with respect to cost, timeliness of results, and respondent burden.

4. APPLICATION OF THE DESIGN MODEL

During the past several years a number of design features have been considered for satisfying the objectives of the NHIS and the Center's other population surveys. In order to systematically evaluate a number of design alternatives, five major design features were identified as best for meeting one or more of the new objectives of the Center's population survey program. The five major design features are listed below along with the design options that were considered for each of the features. A brief discussion of how each feature is related to the objectives of the redesign is also presented.

4.1 Design Features and Options for the Center's Population Surveys

A. Sampling Frame for NHIS

Option 1. Area frame based entirely on geographical areas.

Option 2. List and area frame based primarily on 1980 decennial census information (current design).

B. Multiple Survey Linkage

Option 1. A single integrated survey design for the NHIS, NMCUES, NSFG, and possible NHANES.

Option 2. Independent survey designs for each of the Center's population surveys (current design strategy).

C. Mode of Data Collection for the NHIS

Option 1. A dual mode of data collection using both face-to-face and telephone interviews.

Option 2. A face-to-face mode of data collection (current design).

D. Probability of Selecting Sample Persons in the NHIS

Option 1. Greater probabilities of selecting persons from small domains of special interest (differential probabilities).

Option 2. The same probability of selecting each person in the universe (current design).

E. Rotation of the NHIS Sample

Option 1. Two interviews, one year apart, at each dwelling unit for a subsample of dwelling units.

Option 2. A single interview at each dwelling unit (current cross-sectional design).

From the inception of the NHIS in 1957 the Bureau of the Census has been the data collection agent for NCHS. Since the early 1960's a combined list and area frame have been used for selecting the NHIS sample. The list frame obtained from the decennial census was supplemented by an area frame in non-tracted geographical areas and by a building permit list frame for housing units constructed since the decennial census.

Under the Bureau of the Census's confidentiality provisions in their authorizing legislation they are not allowed to release any information collected in the decennial census. They cannot, therefore, release any of the identifiable person or household information to NCHS in areas where the census list frame was used. Without the identifiable information NCHS would not be able to analyze its results in geographical detail, conduct independant follow-up surveys, use the NHIS sample as a frame for other NCHS surveys, or to link its information to other data sets through geocoding. It is obvious that these capabilities are directly related to the redesign objectives of the NHIS for the 1980's.

One way to avoid the problems of confidentiality is to use an all area sampling frame which is not dependent on the 1980 decennial census. The disadvantage of using an all area frame is the additional cost required to construct the frame and the somewhat larger costs associated with maintaining the sampling frame over the ten year redesign period.

The use of the NHIS sample as a sampling frame for the Center's other population surveys is being considered for linking the Center's

surveys. By using the NHIS sample as a sampling frame for the NSFG and NMCUES the Center could:

a) link the surveys analytically and possibly avoid some duplication of data collection,

b) use the NHIS information to more efficiently oversample the black population for the NSFG, and

c) use the NHIS information to oversample persons with major expenditures in the previous year for the NMCUES.

A possible disadvantage of using the same design for all of its population surveys is the inability to have an optimum design for each survey. The compatability of the survey designs and the feasibility of linking the surveys need to be carefully evaluated.

During the last five years NCHS has invested considerable resources in the study of conducting telephone health interviews. Another paper in this session summarizes the results of the Center's research efforts. In the redesign the use of both face-to-face and telephone interviews is being considered for the NHIS. The main advantages of the telephone interview is its lower cost and closer supervision of interviewers. Before a dual mode of data collection is routinely used in the NHIS, however, a number of methodological questions must be addressed. The redesign telephone interview issues are discussed in a paper by Massey, Marquis, and Tortora [4].

The current NHIS design gives every person in the target population an approximate equal chance of being selected into the NHIS. The oversampling of persons from small domains of the population is being considered to improve the precision of our estimates for the small domains. The most important small domains of special interest in the NHIS are the blacks, the Hispanics, the aged, and persons with low income. The NHIS sample sizes for these domains have often prohibited detailed demographic analysis. The Bureau of the Census has investigated the oversampling of these population domains within the primary sampling units. The effects on the precision of the estimates for these small domains is being compared to losses in precision for the major domains of the population.

In order to improve our ability to measure health trends we have asked the Bureau of the Census to evaluate the partial rotation of the NHIS sample from one year to the next. Under the current cross-sectional design a new sample of dwelling units is interviewed each week. The weekly samples are accumulated over time to produce quarterly, yearly, or bi-yearly estimates. Using the rotation feature a proportion of the NHIS yearly sample of dwelling units would be revisited the following year. By using a proportion of the same dwelling units in consecutive years, the intraclass correlation is increased from one year to the next year and smaller changes from one year to the next can be detected. The disadvantage of using the same dwelling units from year to year is the loss in the precision of our estimates that combine the weekly samples over a two-year period. Since the core questions in the NHIS do not change very often (about every 10 years) the weekly samples can be combined over long periods of time to produce more detailed descriptive statistics for the small domains of the population. A decision on the feature will be based primarily on the effects on the precision of the different types of estimates (estimates of change versus estimates of levels).

4.2 Design Option Sets

After the options had been defined for each of the major design features being considered for the Center's population surveys, the different combinations of the options were enumerated. The 32 possible design option sets for the population survey designs are shown in Table 1. Each option set represents a possible set of specifications for the population survey design. For example, design option set 1 in Table 1 would incorporate all of the proposed new options into the NHIS redesign, whereas design option Set 32 represents the features of the current NHIS design. Before evaluating all 32 design option sets an attempt was made to reduce the number of design option sets to a more manageable level.

4.3 Reduction of Design Option Sets

The initial step taken to reduce the number of design option sets was to determine whether any of the options for the 5 design features could be independently eliminated. Based on discussions within NCHS, it was concluded that the current options for both design features A and B could be eliminated. It was concluded that

- An area frame should be used for the NHIS beginning with the redesign in 1985, and
- An integrated sample design should be used for the NHIS, NMCUES, and NSFG.

A single evaluation was used to help make both of these decisions. In addition to taking into account the advantages associated with the two proposed new design features given in section 4.1, a cost study was conducted. The study included the cost of constructing an all area sampling frame, the cost of maintaining the frame over a ten year period, the cost savings of using the NHIS household information to screen for the NSFG, and the added efficiency of oversampling the NHIS household containing persons with major health problems and large health expenditures for inclusion in the NMCUES. For the analysis it was assumed that the NSFG and the NMCUES would each be conducted twice during the next decade. Under this assumption it was concluded that between 2.2 and 3.7 million dollars (1981 dollars) could be saved by using the NHIS sample as a sampling frame for the NSFG and the NMCUES. This savings would be 2 to 3 times the additional cost required to construct and maintain an all area sampling frame for the

NHIS. It was therefore concluded that an area frame for the NHIS and a linkage of Center's population surveys would both reduce the cost of data collection and improve the analytical utility of the Center's surveys.

The selection of the area frame and the integrated survey options reduces the number of design option sets shown in Table 1 from 32 to 8. The first 8 design option sets in Table 1 represent the remaining designs under consideration. In trying to further reduce the number of design option sets it was decided that the remaining three design features should first be investigated independently. For the mode of data collection feature, the previous telephone research is being evaluated and a Task Force has been formed within NCHS to address both key sample design and questionnaire design methodological issues.

The research for the oversampling of persons from small domains of special interest and the rotation of the NHIS sample from one year to the next has nearly been completed by the Bureau of the Census. It is not clear at this point whether any of the options for these features can be independently selected. Before final design decisions are made for the three remaining design features, the interactions between the features must be considered.

4.4 Interaction Between the Design Features

The final design for the Center's population based surveys will depend a great deal on the compatibility between the various options associated with the design features. In terms of our design model the interactions between the selected options must be evaluated. The first order interactions are of most importance and are shown in Table 2 for the new design options being considered for redesign of the Center's population survey.

A preliminary evaluation of the compatibility between each of the new options is given in Table 2 using a scale from poor to excellent. These ratings assume that the basic designs for the NHIS, NSFG, and NMCUES remain the same as they are currently. The least compatible of the new options appears to be the compatibility between random digit dialed (RDD) telephone interviews in the NHIS and the use of the NHIS as a frame for the NSFG and NMCUES. Both of these latter surveys require face-to-face interviews. If a RDD telephone design were used for the telephone interviews in the NHIS, it would be very inefficient to conduct follow-up face-to-face interviews for NSFG or NMCUES because of the travel costs required to visit the loosely clustered households associated with the RDD design.

The overall efficiency of using the NHIS sample as a frame for the NSFG and NMCUES would depend on the proportion of the NHIS interviews conducted by telephone. Before making a final decision about the incompatibility of these two options more research is required. To make the options more compatible, for example, telephone interviews might be considered for the NMCUES or the NSFG.

Another alternative would be to conduct the telephone interviews in the same geographic areas as the face-to-face NHIS interviews. This might make it more feasible to conduct follow-up interviews for the NSFG or the NMCUES. All of the other first order interactions between the design options appear to be more compatible than the linkage and telephone options. A single statement on compatibility is given for each of the interacting options in Table 2.

For this paper we have assumed that the second and third order interactions are negligible.

Before selecting a final design for the NHIS, there is another set of interactions that must be evaluated. In addition to examining the new design options for the NHIS, a considerable effort has been given to redesigning the basic NHIS sample. The basic sample design features of the NHIS include the number of primary sampling units (PSU's), the size of the secondary sampling units, the total size of the sample, and variables to be used for stratification. It is obvious that the number of PSU's and the size of the ultimate cluster in the NHIS could have a tremendous effect on the efficiency of using the NHIS sample as a sampling frame for the NSFG and the NMCUES.

All of the new features being considered for the Center's population surveys will be effected by the basic design parameters selected for the NHIS. Much of the research being conducted by the Bureau of the Census to determine the optimum design for the NHIS has attempted to take the proposed new design options into account.

The final design will probably require adjustments to both the basic NHIS design parameters and the new options selection for the Center's population surveys. Several papers are being presented by the Bureau of the Census at these meetings on the NHIS redesign.

5.0 SUMMARY

This paper attempts to present an overview of the approach used at NCHS to address the redesign issues associated with the Center's population surveys. Many of the details involved in redesigning a complex survey have been omitted, and a number of less important design features have not been described. The emphasis has been on the process of using a model approach to systematically evaluate the major design features for the redesign.

To the extent possible, the redesign process has been based on obtaining objective measures to evaluate alternative design options. In the end, however, many of the final design decisions will be partially subjective because of the competing survey objectives and program interests. The use of both objective and subjective criteria to design a survey are necessary to produce an end product that is of maximum utility. Singh and Drew [5] present a good discussion of

redesigning a continuous survey in a changing environment in their work on the Canadian Labour Force Survey.

The final design for the Center's population survey has not been selected. The basic design parameters for the NHIS will be selected in the next few months along with decisions on whether to oversample domains of special interest and/or to interview a subset of the NHIS households from one year to the next. Task Forces have been formed within NCHS to determine what research is needed to make decisions on the use of the telephone interview and on the designs and linkages for the NSFG and the NMCUES to the NHIS.

Acknowledgement

We are grateful to Mrs. Phyllis Feldman and Rochelle Weinstein for secretarial assistance.

6. REFERENCES

[1] Plewes, Thomas J. "Planning for the Current Population Survey Redesign Effort," Review of Public Data Use, 8, 1980, pp 141-145.

[2] Burnham, Clinton E. and Massey, James T. "Redesign of the National Health Interview Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings, 1980, pp 115-118.

[3] Sirken, Monroe G. and Royston, Patricia N. "Design Effects in Retrospective Mortality Surveys," Proceedings at the Section on Social Statistics, American Statistical Association Meetings, 1976, pp. 773-777.

[4] Massey, James T., Marquis, Kent H., and Tortora, Robert D. "Methodological Issues Related to Telephone Surveys by Federal Agencies," to be presented at the Section on Social Statistics, American Statistical Association Meetings, 1982.

[5] Singh, M.P. and Drew, J.D. "Redesigning Continuous Surveys in a Changing Environment," Canadian Conference in Applied Statistics, Statistics '81 Canada, Concordia University, Vol. 7, No. 1, 1981.

TABLE 1. ENUMERATION OF THE DESIGN OPTION SETS FOR FIVE DESIGN FEATURES FOR THE POPULATION SURVEY REDESIGN

Design Option Set	Design Feature				
	Sampling Frame	Multiple Survey Linkage	Mode of Data Collection	Selection of Sample Persons	Rotation of Sample
1	Area	Integrated	Dual	Differential	Rotation
2	Area	Integrated	Dual	Differential	No Rotation
3	Area	Integrated	Dual	Uniform	Rotation
4	Area	Integrated	Dual	Uniform	No Rotation
5	Area	Integrated	Face-to-face	Differential	Rotation
6	Area	Integrated	Face-to-face	Differential	No Rotation
7	Area	Integrated	Face-to-face	Uniform	Rotation
8	Area	Integrated	Face-to-face	Uniform	No Rotation
9	Area	Independent	Dual	Differential	Rotation
10	Area	Independent	Dual	Differential	No Rotation
11	Area	Independent	Dual	Uniform	Rotation
12	Area	Independent	Dual	Uniform	No Rotation
13	Area	Independent	Face-to-face	Differential	Rotation
14	Area	Independent	Face-to-face	Differential	No Rotation
15	Area	Independent	Face-to-face	Uniform	Rotation
16	Area	Independent	Face-to-face	Uniform	No Rotation
17	List	Integrated	Dual	Differential	Rotation
18	List	Integrated	Dual	Differential	No Rotation
19	List	Integrated	Dual	Uniform	Rotation
20	List	Integrated	Dual	Uniform	No Rotation
21	List	Integrated	Face-to-face	Differential	Rotation
22	List	Integrated	Face-to-face	Differential	No Rotation
23	List	Integrated	Face-to-face	Uniform	Rotation
24	List	Integrated	Face-to-face	Uniform	No Rotation
25	List	Independent	Dual	Differential	Rotation
26	List	Independent	Dual	Differential	No Rotation
27	List	Independent	Dual	Uniform	Rotation
28	List	Independent	Dual	Uniform	No Rotation
29	List	Independent	Face-to-face	Differential	Rotation
30	List	Independent	Face-to-face	Differential	No Rotation
31	List	Independent	Face-to-face	Uniform	Rotation
32	List	Independent	Face-to-face	Uniform	No Rotation

TABLE 2. TWO-WAY INTERACTIONS FOR NEW DESIGN OPTIONS BEING CONSIDERED FOR POPULATION SURVEY REDESIGN

Interacting Options	Design Compatability
Linkage x Telephone (NHIS)	Poor: NSFG and NMCUES currently require face-to-face interview
Linkage x Rotation	Fair: The NHIS reinterviews reduce yearly sample for NSFG and NMCUES
Linkage x Oversampling	Good: Oversampling of blacks increases yearly sample for NSFG
Telephone x Rotation	Good: The NHIS reinterviews can be conducted by telephone at a lower cost
Telephone x Oversampling	Fair: Screening for special domains using RDD procedures is relatively cheap
Rotation x Oversampling	Good: No apparent adverse effects of interaction

STRATIFYING PRIMARY SAMPLING UNITS WITH THE SAS CLUSTER PROCEDURE

Joe Fred Gonzalez, Jr. and Andrew A. White
Survey Design Staff, Office of Research and Methodology
National Center for Health Statistics

INTRODUCTION

The National Center for Health Statistics (NCHS) is sponsoring the Hispanic Health and Nutrition Examination Survey (HHANES) to be conducted from July 1982 to late 1984. The HHANES, as briefly described by NCHS (1981), uses a multistage stratified probability design to sample Hispanics in selected areas of the United States. The target population for the HHANES is: Hispanics of Mexican descent in selected counties of the five southwestern states (Arizona, California, Colorado, New Mexico, and Texas); Hispanics of Puerto Rican descent in New York City; and Hispanics of Cuban descent in Dade County, Florida. The HHANES within PSU sample design and survey methodology has been pretested in Bronx New York, Miami Florida, and El Paso Texas.

Criteria presented in Massey and Ezzati (1981), such as a minimum number and percent Hispanic in the population of a county or several contiguous counties, were used to define the universe of Primary Sampling Units (PSU's) for the HHANES. The defined universe for the five southwestern states includes 212 counties grouped into 193 PSU's.

For the first stage of the sample design, one PSU per stratum was to be selected with probability proportional to size from a stratified universe of PSU's. Since the variance of an estimate derived from a stratified sample is a function of the within strata component of sampling variance, a major design objective in stratified sampling should be to minimize this component of sampling variance so as to produce more precise estimates. For a multistage design this can be accomplished in part by partitioning the universe of PSU's into optimally homogeneous strata.

Five important PSU characteristics were to be used as the stratification variables: number of Hispanics, percent Hispanic, ratio of the 1980 to the 1970 Hispanic populations, median income, and percent urban. This paper describes the methodology used to stratify the universe of PSU's for the five southwestern states.

Although packaged software cluster analysis techniques such as the SAS routine PROC CLUSTER, described in the SAS User's Guide (1979), address the multivariate problem of grouping vector observations having similar attributes, stratification clustering has usually been implemented with custom written software because the packaged routines do not offer sufficient control over strata characteristics.

A critical constraint for HHANES was that the strata should be of equal Hispanic population size. This was a design requirement for the HHANES since it was operationally desirable to schedule approximately the same number of sample interviews and examinations at each survey location. This requirement can be satisfied by forming equal size strata (clusters) and then applying the same sampling fraction to each stratum.

Because of the time constraints on the survey planning of the HHANES and because the SAS clustering procedure is part of a readily available and widely used software package, this technique was chosen for the HHANES stratification problem. The SAS CLUSTER procedure uses a hierarchical algorithm outlined by Johnson (1967). One major drawback of the algorithm is its inability to impose constraints on the stratum (cluster) sizes. However, by iteratively applying the SAS procedure, the clustering process can be controlled to yield approximately equal size strata.

BRIEF OVERVIEW OF CLUSTER ANALYSIS

Essentially there are two general types of clustering or combining rules in cluster analysis, hierarchical and non-hierarchical. In hierarchical clustering a specific distance measure is defined and the two closest clusters (which could be two single vector observations) are combined into one cluster, then the two closest of the new set of clusters are combined into a single cluster, and so on. There are different types of hierarchical techniques, but perhaps the most popular, as described by Sneath and Sokal (1973), are: single linkage, average linkage, centroid linkage, and complete linkage.

In single linkage the similarity between two clusters is defined as the smallest pairwise distance between vector members of two clusters (the two nearest neighbors of the two clusters). In average linkage the similarity between two clusters is defined as the average pairwise distance between all pairs of vector observations, one from one cluster and one from the other. In centroid linkage the similarity between two clusters is simply defined as the distance between the two cluster mean vectors. Lastly, in complete linkage the similarity between two clusters is defined as the largest pairwise distance between a vector observation in one cluster and a vector observation in the other cluster. Complete linkage is the hierarchical clustering rule that is implemented by the SAS CLUSTER procedure.

The other type of clustering rule is non-hierarchical. Various non-hierarchical techniques are discussed by Anderberg (1973), Friedman and Rubin (1967), Kostanich et al (1981), and Judkins et al (1981). Three specific kinds of non-hierarchical algorithms presented in Friedman and Rubin are: the hill-climbing pass, the forcing pass, and the reassignment pass. In general, non-hierarchical algorithms require far more complicated and time consuming computer programming; so although they may provide more control in stratification, the software has been customized for individual survey use.

APPLICATION OF SAS CLUSTERING ALGORITHM

As mentioned earlier, the clustering algorithm employed in the SAS CLUSTER procedure uses complete linkage as the clustering criterion and hierarchically groups elements or vector observations with similar characteristics. The distance measure is Euclidean and is computed as follows.

Denote a $1 \times k$ row vector by $\underset{\sim}{X}_i' = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and similarly define $\underset{\sim}{X}_j'$. If we let $\underset{\sim}{X}_i'$ denote the i^{th} row vector observation in one cluster and let $\underset{\sim}{X}_j'$ be the j^{th} row vector observation in the other cluster, the distance, d, between the two vector observations is given by:

$$d(\underset{\sim}{X}_i, \underset{\sim}{X}_j) = (\underset{\sim}{X}_i - \underset{\sim}{X}_j)' (\underset{\sim}{X}_i - \underset{\sim}{X}_j).$$

In one, two, or three dimensions, this distance is simply the square of the Euclidean (or linear) distance between the two vector observations.

In the problem being considered, PSU's were characterized by vector observations in five dimensional vector space because there were five stratification variables. As can be seen from the distance equation, the variables or elements of the vector observations with the largest absolute magnitude dominate the distance measure (for example, population totals vs. percentages or ratios). In order to deal with this inequitable contribution from each stratification variable to the distance measure, the STANDARD option within the PROC CLUSTER statement was considered.

The STANDARD option standardizes the distance matrix by dividing each element of the distance matrix by the average distance of all vector observations ($\underset{\sim}{X}_i$) from the overall mean vector ($\underset{\sim}{\bar{X}}$). That is, each element of distance matrix is divided by $\sum_{i=1}^{n} [d(\underset{\sim}{X}_i, \underset{\sim}{\bar{X}})] / n$, where n is the number of vector observations.

As can be seen from the above discussion, this type of standardization does not solve the problem of unequal contribution to the distance measure by the stratification variables.

Ranking is one easy way to equalize the relative contribution to the distance measure of the stratification variables and to assign equal importance to each variable by transforming the vectors of observed values to vectors of rank values. In this case, PSU's were ranked on each of the five variables by applying the SAS PROC RANK procedure. For those cases where tied values appeared, the mean rank of the corresponding ranks was assigned. Then the SAS Cluster procedure was applied to the vectors of ranks of the PSU's.

Due to operational and cost considerations it was planned that the HHANES be conducted at 14 different locations (cities) in the Southwest. Since the 1980 Hispanic populations for the PSU's for the PSU's were used as measures of size (MOS's) in determining the relative sampling probabilities, the total Hispanic population for the 193 PSU's was divided by 14 to obtain the desired stratum size. Because of their large Hispanic populations, Los Angeles, California and Bexar County (San Antonio), Texas became certainty strata. This left a remainder of 191 PSU's to be clustered into 12 non-certainty strata.

The first few non-certainty strata were determined "manually" by using the 193 x 193 pairwise distance matrix produced by the SAS clustering algorithm. This matrix was extremely useful since it gave the distance measures between all PSU's in the universe. The "manual" stratification was intended to pair similar PSU's with a substantial number of Hispanics (approximately 150,000 or more).

Beginning with the formation of the certainty strata, after each successive stratum was identified, it was withdrawn from the universe without replacement. Then a new stratum cutoff size was recomputed by dividing the remaining Hispanic population in the universe by the number of strata yet to be formed. In practice, it was almost impossible to achieve the same size for each stratum, therefore, a tolerance level of ±10% from the cutoff value was allowed for the non-certainty strata.

The remainder of the strata (clusters) were formed using the SAS clustering algorithm. The algorithm was constrained to group the remaining PSU's into the number of strata (clusters) still needed to be formed. The SAS output included a list of all PSU's remaining in the universe to be clustered along with their respective vectors of raw variables and vectors of the rank variables. Additional output included: various measures used in the Cluster Analysis, a distance matrix for all PSU's remaining in the universe, a cluster assignment listing for the number of clusters requested, and a cluster map showing the hierarchical path followed in producing the desired clusters. Due to physical limitations of computer printouts, the output for the cluster map had to be cut and pasted together in order to get a graphic view of the entire hierarchical clustering process.

Figure 1 demonstrates a fictitious example of a cluster map. Each full column of asterisks on the map represents a PSU. Down the far left side of the map, the number of clusters appears for each hierarchical level of clustering. The uppermost row on the map shows the initial hierarchical level of clustering where each PSU's vector of ranks is treated as a single cluster. The lowest row on the map represents one cluster of all the PSU's remaining in the universe. Across the top of the map are the names of the PSU's and their corresponding Hispanic populations. Actually two separate runs of the cluster map were produced - one with the names of the PSU's and the other with the PSU Hispanic populations as labels. Then the Hispanic population labels were cut and pasted above their respective PSU name labels.

As the cluster map is read downward and across, one can see how adjacent columns (PSU's) were collapsed hierarchically to form clusters.

Figure 1 - Example Cluster Map

Number of Clusters	PSU 1 N_1	PSU 2 N_2	PSU 3 N_3	PSU 4 N_4	PSU 5 N_5	PSU 6 N_6	PSU 7 N_7	PSU 8 N_8	PSU 9 N_9	PSU 10 N_{10}
10	*	*	*	*	*	*	*	*	*	*
9	*	*	*	*	*	*	*	****	*	
8	*	*	****	*	*	*	****	*		
7	*	*******	*	*	*	****	*			
6	*	*******	****	*	****	*				
5	*	*******	****	*******	*					
4	**********	****	*******	*						
3	**********	****	**********							
2	**********	**********								
1	**********									

Note: N_i denotes the Hispanic population of the i^{th} PSU.

Candidate clusters were screened and the cluster that amalgamated the earliest (i.e., had smallest distance measures or were "closest" together) and achieved ±10% of the current stratum cutoff size was selected. If all of the constrained clusters had populations that were too large, the sub-cluster which formed first within the size constraint was selected as a stratum.

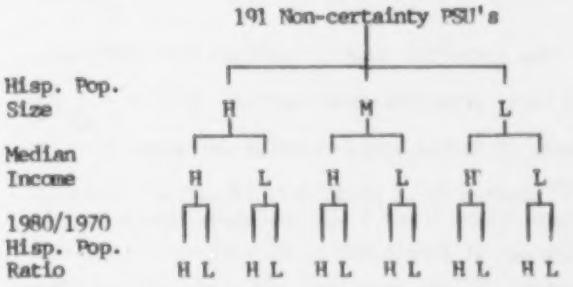
Sometimes several subclusters had to be evaluated before particular ones were combined to achieve the stratum cutoff size. In one instance, a candidate cluster was considered, but it fell short of the desired population cutoff. However, a nearby pair of PSU's could provide the additional population needed. Also, there was a nearby single PSU which could likewise provide the needed population. In order to decide whether to combine the pair of PSU's or the single PSU with the candidate cluster, the distance between the mean vector of ranks of the candidate cluster and the vector of ranks of the pair of PSU's was computed. This distance was then compared with the distance between the mean vector of ranks of the candidate cluster and the vector of ranks of the single PSU. The first distance was smaller, therefore, the pair of PSU's was added to the candidate cluster and a stratum was identified.

After the first formed cluster was identified, the PSU's comprising that cluster were removed from the universe and the stratum cutoff size was recomputed for the next stratum to be formed from the succeeding SAS computer run. Prior to invoking the SAS clustering algorithm on each successive run, the SAS PROC RANK routine was used to rerank the raw variables of the remaining PSU's in the universe. The above procedures were repeated until all of the PSU's were stratified.

COMPARISON WITH A CONVENTIONAL APPROACH

In order to measure the relative efficiency of the clustering stratification scheme, the 191 non-certainty PSU's were grouped into 12 strata using a conventional approach to first stage stratification. This approach involved successively sorting (via the SAS PROC SORT routine) the non-certainty PSU's by three stratification variables: number of Hispanics, median income, and the ratio of the 1980 to the 1970 PSU Hispanic populations. Because the universe was to be divided into only twelve strata, the sorting method could not use all five stratification variables used in the clustering method. Figure 2 illustrates this sorting procedure in schematic form.

Figure 2 - Sorting Method of Stratification



After ordering the PSU's by number of Hispanics, the entire set of non-certainty PSU's was subdivided into three groups with approximately equal Hispanic populations. These three groups were made up of PSU's with High (H), Medium (M), and Low (L) Hispanic population sizes, respectively. By ordering the PSU's by median income, each of the three groups of PSU's was then divided into two subgroups with approximately equal Hispanic population size, composed of High (H) and Low (L) median income PSU's. This resulted in six subgroups of approximately equal Hispanic population size. Finally the PSU's within each of the six subgroups were ordered by the ratio of the 1980 to the 1970 Hispanic populations and again divided into High (H) and (L) strata with approximately equal Hispanic populations. Table 1 shows the number of PSU's and the Hispanic population by stratum for strata formed by both the clustering and sorting techniques.

The purpose of stratification prior to sampling is to minimize the within strata variation with respect to measures correlated with the stratification variables. Actual estimates of survey variables are usually unavailable during the planning stage, therefore, the efficiency of the stratification methods was assessed with respect to the stratification variables.

Table 1 - Strata Characteristics by Stratification Method

Stratum Number (h)	Clustering		Sorting	
	Number of PSU's	Hispanic Pop.	Number of PSU's	Hispanic Pop.
1*	1	2,065,727	1	2,065,727
2†	1	460,911	1	460,911
3	2	425,610	2	364,294
4	2	561,507	2	501,787
5	3	479,816	2	527,213
6	3	493,459	2	655,406
7	4	514,450	4	396,235
8	4	532,890	4	534,545
9	7	457,644	5	488,563
10	7	489,476	5	534,031
11	12	545,812	23	492,186
12	32	544,237	25	479,805
13	40	454,839	48	493,308
14	75	458,653	69	491,020

* Certainty stratum: Los Angeles, CA.
† Certainty stratum: Bexar Co.(San Antonio), TX.

The criterion used to evaluate the efficiency of both stratification methods was $W = \sum_{h=1}^{14} S_h^2$, where S_h^2 is the within stratum variance for the h^{th} stratum for a given stratification variable. Since strata 1 and 2 are certainty strata composed of single PSU's, $S_1^2 = S_2^2 = 0$. The criterion, W, was computed for each of the five stratification variables for both the clustering and the sorting methods; these values of W are displayed in Table 2.

Table 2 - W Criterion by Stratification Method and Variable

Variable	Clustering (1)	Sorting (2)	Ratio* (3)
Hisp. Pop.	6.25×10^{10}	1.11×10^{10}	5.61
Percent Hisp.	1.58×10^{5}	2.45×10^{5}	0.64
1980/1970 Ratio	2.71×10^{4}	3.69×10^{4}	0.74
Median Income	1.74×10^{7}	2.08×10^{7}	0.84
Percent Urban	2.80×10^{3}	3.36×10^{3}	0.83

* Ratio of column (1) to column (2).

As can be seen from Table 2, the clustering method used in this paper produced smaller W values for all of the stratification variables except Hispanic population. This is due to the fact that Hispanic population was used as the primary ordering variable in the sorting method, whereas all five of the stratification variables were assigned equal importance in the clustering method. In fact, with the clustering method, a vector of arbitrary weights can be assigned to the stratification variables to reflect their expected relative importance in reducing the variance of specific estimates. Because the HHANES is a general purpose survey intended to provide estimates of many varied characteristics, the SAS CLUSTER procedure was allowed to default to a unit weight vector in this application.

There are criteria other than W that may be used to further investigate the characteristics of a universe stratified by more than one technique. For example, the actual Euclidean distance measure used to determine clusters could be computed for the strata as formed by the sorting method. Another factor to be considered is the amount of resources required to implement each method. For the problem addressed in this paper, both methods required approximately one person-week to complete the stratification; however, the clustering method required far more computer time (31.2 CPU minutes as opposed to 6.5 CPU minutes for the sorting method). For organizations with a dedicated computer, this makes little or no difference, but if computer access is purchased on a CPU time basis, the cost factor may be an important consideration in the selection of a stratification method.

SUMMARY

A standard cluster analysis technique, implemented by the SAS routine PROC CLUSTER, was successfully used to stratify the PSU's in the HHANES universe prior to first stage sample selection. A critical constraint for HHANES was that the strata should be of equal Hispanic population size. By iteratively applying the SAS procedure to a vector of ranks of the five stratification variables, it was possible to assign equal importance to each of the stratification variables and to stratify (cluster) the HHANES PSU's into homogeneous groups of approximately equal size.

The universe was also stratified using a conventional sorting method and the efficiency of the resulting stratification was compared with that resulting from clustering with the SAS routine. Whereas with the clustering method it was possible to use five stratification variables, with the sorting method the number of variables was constrained to three (as a function of the number of strata and the number of levels into which each variable was categorized). Also, clustering enabled the stratification variables to be assigned equal importance, whereas sorting placed varying importance on the variables according to sort order.

The clustering method was found to produce lower within strata variances for four of the five stratification variables used. The sorting method produced a lower within strata variance for the primary sorting variable.

The SAS CLUSTER procedure was a viable multivariate method of stratification as described in this paper. It should prove to be an equally valuable tool for similar surveys.

Acknowledgements

We are grateful to James T. Massey and Robert J. Casady for their suggestions and guidance during this project. The methodology discussed in this paper required that large quantities of SAS computer output be carefully reviewed; we wish to thank Ms. Trena Ezzati for her help during this stage of the project. We are also grateful to Mrs. Phyllis Feldman for secretarial assistance during the preparation of this manuscript.

REFERENCES

1. Anderberg, M.R. (1973). **Cluster Analysis for Applications.** Academic Press, New York.

2. Friedman, H.P. and Rubin, J. (1967). On Some Invariant Criteria for Grouping Data. **Journal of the American Statistical Association 62,** 1159-1178.

3. Johnson, Stephen C. (1967). Hierarchical Clustering Schemes, **Psychometrica, XXXII,** pp. 241-254.

4. Judkins, D.R. and Singh, R.P. (1981). Using Clustering Algorithms to Stratify Primary Sampling Units, **Proc. Sect. Surv. Res. Meth., Amer. Stat. Assoc.,** pp. 279-284.

5. Kostanich, D.L., et al (1981). Modification of Friedman - Rubin's Clustering Algorithm for Use in Stratified PPS Sampling, **Proc. Sect. Surv. Res. Meth., Amer. Stat. Assoc.,** pp. 285-290.

6. Massey, J.T. and Ezzati, T.M. (1981). HHANES Universe and Estimated Coverage, unpublished memorandum, National Center for Health Statistics, August 7.

7. National Center for Health Statistics: The Hispanic Health and Nutrition Examination Survey-Fact Sheet. DHHS Publication No. (PHS) 81-1287. Public Health Service, Washington, D.C. U.S. Government Printing Office, Sept. 1981.

8. SAS User's Guide (1979) Edition, SAS Institute Inc., Cary, North Carolina.

9. Sneath, P.H.A. and Sokal, R.R. (1973). **Numerical Taxonomy,** W.H. Freeman and Co., San Francisco, pp. 216-228.

NCHS RESEARCH ON THE TELEPHONE INTERVIEW: SOME OBSERVATIONS

Owen T. Thornberry, Jr., and Gail S. Poe

I. Introduction

The National Center for Health Statistics (NCHS) interest in the telephone interview as a mode of data collection began in 1976. That year the Center Director established a committee to assess and report on the potential applicability of the telephone interview and telephone survey methodology to the data collection needs of NCHS.

This interest was motivated by a number of factors, including:

1. The increasing expense of the traditional personal interview,
2. Some research suggesting comparable results from the personal and telephone modes,
3. The potential for expansion of the Center's methodological research base, including pretesting of supplements to the National Health Interview Survey (NHIS),
4. The opportunity to provide rapid reporting on topics of interest within the Public Health Service,
5. The opportunity to be more responsive to State and local data needs and other demands for small area data,
6. The potential for the development of an in-house data collection system within NCHS, and
7. The flexibility and quality control inherent in a centralized data collection system.

II. Overview of Projects

A number of projects grew out of the deliberations of the NCHS work group.

1. Based on data from the NHIS, an analysis of the characteristics of telephone and non-telephone households and procedures for correcting for undercoverage bias in random-digit-dialed national health surveys (1).
2. The development of an in-house National Telephone Health Interview System (NTHIS) at NCHS (1-14).
3. A major contract with the Survey Research Center of the University of Michigan for an evaluation of the telephone and personal interview for the collection of National Health Interview Survey data and for methodological research on the telephone interview (15-27).
4. A national random-digit-dialed (RDD) telephone survey of personal health practices sponsored by the Division of Analysis of NCHS (28, 29).
5. Several efforts aimed at data collection at the State and local level, including:
   (a) Two contracts from the Division of the Cooperative Health Statistics System of NCHS for local area RDD surveys, methodological research on the telephone interview and the refinement of small area estimation techniques (30, 31, 33-35).
   (b) A contract between the Office of Research and Methodology and UCLA for the development of a manual for small area surveys using random digit dialing (32), and
   (c) An Applied Statistics Training Institute course on health interview surveys with a focus on the State or local area telephone survey.
6. The use of the telephone interview for two of the five household interviews in the longitudinal National Medical Care Utilization and Expenditure Survey (NMCUES) sponsored by NCHS and the Health Care Financing Administration. This complex study included both a national sample of households and a 4-State sample of Medicaid households. In those rounds where the telephone interview was used, around 85 percent for the national sample and 60 percent for the Medicaid sample were completed by telephone. The methodological findings should provide valuable insights relative to the use of the telephone for interviews or reinterviews of low income households.
7. The telephone was used in a pretest of a proposed supplement to the NHIS on alcohol consumption and health practices.
8. Staff of NCHS also have provided technical assistance in conducting telephone surveys to a number of federal and State and local agencies.

III. The NCHS National Telephone Health Interview System (NTHIS)

The NTHIS, initiated at NCHS in 1978, evolved through three stages: developmental work, methodological research, and finally substantive data collection. The first survey, partially funded by the Office on Smoking and Health, was conducted during 1979 for developmental and methodological purposes, including a comparison

of data gathered in the National RDD survey with comparable data from the NHIS during the same time period and experimentation on the placement of household roster questions and procedures for asking family income and other sociodemographic questions. Papers documenting this research have been presented at previous ASA meetings (2-5, 7, 10).

The second survey, the National Longitudinal Telephone Health Survey on Cigarette Smoking Practices was conducted in 1980 and 1981. This survey was designed to include three contacts with each respondent, at six-month intervals, to track changes in smoking behavior over a one-year period. Unfortunately, because of budget cuts, this project was terminated after two waves of interviewing.

All adult household members in this survey were asked to respond for themselves. A Wave I household response rate of 84 percent and a person response rate of 73 percent were obtained. Locator information was obtained during the Wave I interview and two letters and a postcard for address changes were mailed prior to Wave II. Around 95 percent of the Wave I respondent households were recontacted for Wave II. The Wave II household response rate was 88 percent. The analysis of data from this survey is underway and draft reports should be available later in the year.

IV. University of Michigan Contract[1/]

The NCHS contract with the Survey Research Center (SRC) of the University of Michigan had two primary purposes: first, to compare and evaluate face-to-face and telephone interview methods for collecting data in the National Health Interview Survey, and second, to conduct a number of experiments specific to a national random-digit-dialed telephone survey. A draft of the final report from this contract is under review and will be published later in the year by NCHS as a Series 2 report. Various aspects of the research have been presented at previous ASA annual meetings (22-25).

In the fourth quarter of 1979, SRC conducted a national probability RDD telephone survey of about 4,300 families--yielding data on 8,200 persons 17 years or older--using a modified NHIS questionnaire. At the same time, Bureau of the Census contacted a separate national sample of families as part of the ongoing NHIS--yielding data for 19,800 persons 17 years and older.

In the SRC telephone survey one person from each family responded for all adult family members. The research design included the random assignment of a set of treatments to sample telephone numbers as is shown in Table I. The net result was three experimental groups.

[1/] The discussion provided below was taken in part from the contract final report and from various papers prepared by SRC (15-27).

(a) First, families were assigned to one of two interviewing methods. The control version specified a behavior on the part of the interviewer that was similar to that of Bureau of the Census interviews taped before the survey. The experimental version used designated instructions and feedback to the respondents and also sought a commitment from the respondent to answer carefully and honestly.

(b) Second, two alternative respondent rules were used. In the knowledgeable adult half sample any adult answering the telephone who judged themselves capable of answering the health questions responded for all adults in the family. In the random respondent sample, a household listing was taken from the phone answerer. One respondent 17 years or older was then selected to respond for all adults.

(c) In the third experimental treatment, half samples of telephone numbers were assigned to either a CATI or paper and pencil version of the questionnaire.

In addition to these three basic experimental treatments, the design included randomized assignment of sample numbers to interviewers, an experiment on response rates that used dual contacts on some sample households, and the simultaneous monitoring of the interview by persons who recorded the behavior of the interviewers.

For the telephone/personal interview comparison, the telephone response rate was well below that for the regular NHIS - 80 percent versus 95 percent. Table 2 provides estimated response rates for various subgroups of the knowledgeable adult and the random respondent samples. Based on the assumption of no nonresponse bias in the NHIS data, response rates for subgroups within the telephone sample can be estimated by the product of the overall telephone response rate and the ratio of the proportion of persons in the subclass for the telephone survey to the same proportion for NHIS. (P_i,SRC/P_i,HIS=R_{SRC})

As shown, for the knowledgeable adult rule there are low response rates for the elderly (66 percent versus 80 percent overall) and the least educated (74 percent versus 80 percent) and no differences by sex or race. Note that interpretation of income differences are difficult given relatively high item nonresponse in both the telephone survey and the NHIS. The response rates for the random respondent group are generally lower for all subgroups. Response rates are lower for both the very young and elderly. Further, females have higher response rates than males and whites higher than nonwhites.

Regarding the substantive comparison between the telephone and personal interview data, there is a tendency for telephone respondents to report more health events than respondents in the NHIS personal interview. See for example Table 3. In

a comparison on 64 statistics, telephone respondents evidenced lower reporting than NHIS respondents on only 12.

Additional analyses were performed by SRC to determine if subgroups of the population exhibited variation in the differences between modes. Three demographic variables were used--age, sex, and years of education. The demographic variables were chosen because they were found to be correlates of health status or of likelihood of cooperation with the survey. Log linear models were fitted to a variety of tables. All of the models exhibited the significant effect of higher telephone survey reporting of health events. That is, in summary of the multivariate models using mode and a demographic variable as predictors of the health variables, it was found that the higher levels of reporting among telephone respondents appeared to remain within different sex, age, and education groups. The only exception is a tendency for older respondents to report more in face-to-face interviews.

As noted previously, the telephone sample cases were randomly assigned to one of two interviewing treatments. The control procedure was an attempt to approximate a Census interview--i.e., to standardize interviewer behavior as much as possible across modes. This approach restricted the interviewer to asking questions as worded in the questionnaire and to using specified probes and introductory statements.

For comparison to this procedure, an experimental treatment was administered to the other half of the telephone sample. This procedure incorporated three techniques in the NHIS questionnaire: commitment, instructions, and feedback. Commitment involved a verbal agreement by the respondent to give accurate and complete information. Instructions were in the form of statements in the questionnaire at various points for the interviewer to read--e.g., "This is sometimes hard to remember, so please take your time." Feedback was both positive ("I see; this is the kind of exact answer we need") and negative ("You answered that quickly. Are there any days you might have overlooked?").

For almost all health events, there were higher levels of reporting for the experimental groups. See Table 4 for example. Further analysis was performed on different demographic subclasses in order to search for interaction effects involving the experimental interviewing methods. Age, sex, and education were used as predictor variables. The effects of the experimental treatment were not reduced when controls for respondent characteristics were applied. The experimental techniques appeared to facilitate accurate reporting on health variables in this study.

Regarding proxy reporting, it is generally assumed that self-reporting is more accurate. Because of randomized selection, the random respondent half sample provides a test of the effects of self-reporting. Given equal expected values for the randomly selected respondent and for other family members, any differences between the statistics on the two must arise from nonresponse or response error.

As shown in Table 5, the findings from this study are in sharp contrast to earlier studies that did not employ a random assignment to the self or proxy reporting condition. Over several different types of health measures, more health events tend to be reported for the proxied cases than for the respondents themselves. SRC proposes a number of possible hypotheses to explain this unexpected finding (15, 23). In any case the results are contrary to dominant beliefs and replication and experimentation is called for.

As noted, the design also included a random assignment of each family to either a CATI or non-CATI questionnaire condition. There was the possibility that differences might arise in questions that had complicated skip patterns and that some other questions might differ in terms of item completion rates. In general, these differences did not appear. A variety of measures on the total sample and on subclasses showed no differences between CATI and non-CATI.

As a final point, it was found that interviewer variance components for this study were substantially lower than those of similar studies in the past. This was probably a result of the specified interviewer behavior, training, monitoring, and feedback procedures of this project.

V. Observations

There are a number of observations which can be made based on the NCHS conducted or sponsored research. The discussion which follows is not intended to be all inclusive but to highlight some of the significant issues relative to the use of the telephone interview in future NCHS data collection.

1. Undercoverage is a problem that can be dealt with. Most households have telephones (94 percent) and there exists a considerable amount of information on the characteristics of non-telephone households. While telephone surveys of the low income segment of the population remain problematic, adjustments for the error introduced by the exclusion of non-telephone households in most RDD surveys can be made through post-survey adjustments and/or dual-frame designs. In order to obtain data on access to telephones and to more efficiently design RDD surveys, the NCHS and Census have developed the questions in Figure 1 for inclusion in the third and fourth quarter of the NHIS this year.

2. A more serious problem is the apparent lower response rates overall and for certain population subgroups with RDD surveys in contrast to those conducted by

personal interview--or at least the face-to-face surveys conducted by the Bureau of the Census. Table 6 in the attachment provides estimated response rates from NCHS sponsored RDD surveys. Household response rates range between 76-86 percent and person rates somewhat lower. This of course is well below Census rates of around 95 percent. It is hoped that Census RDD pilot studies will achieve response rates in the neighborhood of 90 percent.

3. Regarding population subgroups, we need to develop procedures or techniques for obtaining cooperation from the low income and low education segments of the population and from the elderly. The elderly have a higher proportion with telephone coverage than any other age group but also have the highest rate of nonresponse. Further, the SRC research suggests possible better reporting for the elderly in person rather than over the telephone; the only age group with this finding. As SRC has noted, this should cause some concern for researchers interested in health variables because of the ubiquitous correlation of age and health status.

4. Regarding item nonresponse, the reporting of income appears much poorer over the telephone. Further, research by Monsees and Massey (2) suggests that the information provided is highly sensitive to the way the question is administered. This item and others where response is facilitated in the personal interview through the use of a "flash card" require much experimentation and developmental work for satisfactory telephone application (although it is worth noting that a high nonresponse on income remains a problem with the personal interview, even for Census).

5. Apparently telephone interviews can be considerably longer than many originally assumed. The SRC interview and others exceeded thirty minutes on the average with no apparent impact on response rates or data quality. Almost all breakoffs in telephone surveys occur in the first minute or so. If we can learn how to keep potential respondents on for a couple of minutes, response rates should increase significantly.

6. There appears to be a tendency on the part of both the interviewer and the respondent to conduct a telephone interview at a very rapid pace. The SRC findings relative to the experimental interviewing techniques suggest that this is a mistake which probably results in less reporting and lower quality data.

7. The SRC findings on self versus proxy reporting are important because of the implications for response rates and cost. This represents an important area for "validation" studies.

8. In most respects the data obtained in an RDD telephone survey appears to be comparable to that obtained with the personal interview. Indeed, data from the SRC survey suggests the possibility of improved reporting over the telephone. In fact, it is quite probable that the issue really is not one of differences between modes; rather it is one of the overall quality of survey design--in particular data collection procedures--and one of taking advantage of the strengths of the telephone approach. For example, SRC suspects the following as contributing to the apparent improved reporting of health events in the telephone survey.

   (1) The rigorous training of interviewers preceding the survey.

   (2) The close contact between the principal investigators and the interviewing staff.

   (3) Consequently, the high morale among the interviewers.

   (4) The greater specification of the interviewing task through adjustment of the questionnaire to include specific instructions about interviewer feedback, and probing.

   (5) Continual monitoring of interviews in progress, with feedback given to interviewers weekly on their performance.

   (6) Supervisory review of all cases after completion.

   (7) The sole attention of the interviewer given to a single respondent in the household instead of group interviews as in NHIS.

   (8) Certain differences in the nature of proxy reporting in the two modes. For example, the NHIS interview asks the same question about every family member in sequence, one by one until the whole questionnaire is complete. The telephone interview asks the entire questionnaire for each family member singly.

   These findings suggest that it is most important to take full advantage of the strengths of a centralized telephone system--the ability to monitor, standardize interviewer behavior, etc. In this respect, the SRC findings regarding the manipulation of interviewer behavior through experimental interviewing techniques (commitment, instructions, feedback) are very important. This also has numerous

implications relative to administrative structures and to CATI systems.

VI. Future NCHS Research

Because of lack of resources, the Center management decided to terminate the NTHIS last December.

However, NCHS will remain very active in several areas:

(1) The use of a dual-frame (telephone/personal interview) approach is under consideration for the redesign of NCHS surveys.

(2) NCHS staff are working closely with Census in the development of their RDD and CATI systems, in particular relative to potential application in the NHIS.

(3) There is the possibility of funding some research through one percent evaluation funds.

(4) The Center will continue to provide technical support to federal, State, and local agencies.

(5) The analysis of data already collected will continue, and

(6) The telephone will no doubt be used for the pretesting of some proposed NHIS supplements for future years.

Reports Based on NCHS Research on Telephone Interview

I. Reports Based on NCHS In-house Research

1. Thornberry, Massey: Correcting for undercoverage bias in random digit dialed national health surveys (1978, ASA)

2. Monsees, Massey: Adapting procedures for collecting demographic data in a personal interview to a telephone interview (1979, ASA)

3. Bercini, Massey: Obtaining the household roster in a telephone survey: the impact of names and placement on response rates (1979, ASA)

4. Fitti: Some results from the telephone health interview system (1979, ASA)

5. Massey, Barker, Moss: Comparative results of face-to-face and telephone interviews in a survey of cigarette smoking (1979, APHA)

6. Casady, Sirken: A multiplicity estimator for multiple frame sampling (1980, ASA)

7. Massey, Barker, Hsiung: An investigation of response in a telephone survey (1981, ASA)

8. Casady, Snowden, Sirken: A study of dual frame estimators for the National Health Interview Survey (1981, ASA)

9. Cannell, Thornberry, Fuchsberg: Research on data collection methods--the National Health Interview Survey (1981, ASA)

10. Botman, Massey, Shimizu: Effects of weighting adjustments on estimates from a random digit dialed survey (to be presented at the 1982 ASA annual meeting)

11. Thornberry, Nicholls, Kulpinski: An assessment of the use of the telephone interview in Federal surveys (to be presented at the 1982 ASA annual meeting)

12. Massey, Marquis, Tortora: Methodological issues related to telephone surveys by Federal agencies (to be presented at the 1982 ASA annual meeting)

13. Massey, Fitti: Classifying drinkers from a telephone survey using two different reference periods (to be presented at the 110th Annual Meeting of American Public Health Association, 1982)

14. Thornberry, Cannell, Groves, Miller: Research on the telephone interview - unresolved issues (to be presented at the 110th Annual Meeting of the American Public Health Association, 1982)

II. Reports Based on NCHS Contract No. 233-78-2034 with the Survey Research Center, University of Michigan "An Evaluation of the Telephone and Personal Interview as Mechanisms for the Collection of Health Interview Data"

15. Cannell, Groves, Magilavy, Mathiowetz, Miller: An experimental comparison of telephone and personal health surveys, 2 vols., final report under NCHS Contract No. 233-78-2034, March 1982 (To be published by NCHS)

16. Survey Research Center, University of Michigan: Technical, administrative training and supervisory procedures for data collection under NCHS Contract No. 233-78-2034

17. Groves, Magilavy, Bialecki: Report on sampling activities: telephone sample design and evaluation of matching exchanges to Census-based units (July 1980 - contract report to NCHS)

18. Groves, Magilavy: Initial estimates of interviewer effects (September 1980 - contract report to NCHS)

19. Cannell, Groves: Relationships between relative bias and interviewer variance components associated with the experimental interviewing behaviors (September 1981 - contract report to NCHS)

20. Cannell, Groves: Comparison of results between computer assisted telephone interviewing (CATI) and paper and pencil questionnaires (September 1981 contract report to NCHS)

21. Cannell, Groves: The effects of postsurvey adjustments on estimates from the Health Interview Survey and the SRC Telelphone Survey (September 1981 - contract report to NCHS)

22. Groves, Magilavy, Mathiowetz: The process of interviewer variability: evidence from telephone surveys (August 1981, ASA)

23. Groves, Mathiowetz: An experiment to measure the effects of respondent rules on health survey responses (August 1981, ASA)

24. Cannell, Groves, Miller: The effects of mode of data collection on health survey data (August 1981, ASA)

25. Miller: Alternative scaling procedures for attitude measurement in telephone surveys (August 1981, ASA)

26. Groves, Magilavy: Increasing response rates to telephone surveys: a door in the face for foot-in-the door (POQ, Fall 1981)

27. Miller: A comparison of telephone and personal interviews in the Health Interview Survey (Fourth Conference on Health Survey Research Methods, 1982)

III. Reports Based on Other NCHS Contract Telephone Surveys

28. Danchik, Schoenborn, Elinson: Highlights from Wave I of the national survey of personal health practices and consequences: U.S., 1979, NCHS, Series 15 - No. 1 (DHHS Pub. No. PHS-81-1162)

29. Schoenborn, Danchik, Elinson: Basic data from Wave I of the national survey of personal health practices and consequences: U.S., 1979, NCHS, Series 15 - No. 2 (DHHS Pub. No. PHS-81-1163)

30. Health Services Research and Development Center, Johns Hopkins Medical Institutions: Evaluation of small area estimation techniques for health planning needs in a local setting, final report under NCHS Contract No. 233-78-2052, January 1981

31. DiGaetano, MacKenzie, Waksberg, Yaffe: Synthetic estimates for local areas from the health interview survey (1980, ASA)

32. Institute for Social Science Research, UCLA: A manual for small area surveys using random digit dialing, final report under NCHS contract, January 1980

33. Weeks, Kulka, Lessler, Whitmore: A comparison of the telephone interview mode and face-to-face interview mode for conducting local household health surveys (Fourth Conference on Health Survey Research Methods, 1982)

34. Weeks, Kulka, Lessler, Whitmore: A comparison of face-to-face and telephone modes for collecting health data from households (37th Annual Conference of the American Association for Public Opinion Research, 1982)

35. Lessler, Kulka: The impact of differences in survey interviewing modes on policy decisions for the aged (to be presented at the 110th Annual Meeting of the American Public Health Association, 1982)

Table 1. Number of Persons with Interview Data by Experimental Groups[1/]

Mode	Random Adult Respondent Rule			Knowledgeable Adult Respondent Rule			Total
	Experimental Interviewing Behavior	Control Interviewing Behavior	Total	Experimental Interviewing Behavior	Control Interviewing Behavior	Total	
CATI	837	933	1,770	964	937	1,901	3,671
NonCATI	1,009	1,095	2,104	1,064	1,162	2,226	4,330
Total	1,846	2,028	3,874	2,028	2,099	4,127	8,001

[1/] 209 individuals in households where the random respondent could not be interviewed, but where some other family member responded have been removed from this table.

Table 2. Estimated Response Rates for Selected Population Subgroups

Population Subgroup	Estimated Response Rate	
	Knowledgeable Adult Respondent Rule	Random Adult Respondent Rule
	Percent	
Sex		
Male	82.6%	72.2%
Female	80.7	79.1
Age		
17-24	82.0	61.5
25-44	87.8	85.6
45-64	81.2	81.3
65-74	66.3	56.7
Race		
White	81.5	76.5
Nonwhite	82.3	68.7
Education		
0-11 years	74.1	58.4
12 years	79.6	72.4
13+ years	91.2	95.6
Income		
Less than $10,000	73.3	75.0
$10,000-24,999	93.6	85.5
$25,000 or more	71.4	61.5
Marital Status		
Married	81.4	78.7
Widowed	75.8	58.5
Divorced	88.8	96.5
Separated	79.9	86.1
Never Married	82.8	65.1
Usual Activity		
Working	82.5	80.3
Keeping House	82.9	84.9
Other	76.8	47.2

Table 3. Percentage of Persons with Selected Health Characteristics by Mode of Data Collection and Sex of Sample Person

Selected Health Characteristic	Male		Female	
	NHIS Data	Total Telephone Sample	NHIS Data	Total Telephone Sample
At least one work loss day (two weeks)	5.1	8.1*	4.2	7.2*
At least one cut down day (two weeks)	7.1	8.7*	7.9	10.7*
At least one doctor visit (two weeks)	11.3	13.6*	15.4	17.9*
At least one dental visit (two weeks)	5.4	6.9*	5.3	7.4*
At least one bed day (twelve months)	42.3	50.3*	49.4	55.6*
Excellent subjective health status	52.3	55.8*	59.1	59.5
Approximation	8400	3800	10000	4400

*Telephone percentage significantly different from NHIS percentage at .05 level.

Table 4. Percentage of Persons with Selected Health Characteristics by Experimental Interview Treatment

Selected Health Characteristic	Control Form	Experimental Form
Percentage with at least one: (two week recall)		
Bed day	7.3	10.0
Work loss day	6.3	8.8
Cut down day	8.4	11.5
Dental visit	6.8	7.4
Doctor visit	17.4	17.5
Acute condition	14.9	17.7
Percentage with at least one: (one year recall)		
Chronic condition	29.2	35.8
Limitation of activity	20.4	27.6
Doctor visit	72.6	74.5
Health status (% excellent)	42.1	41.3
Hospitalization	13.4	12.5
Percentage in modal category:		
12 month bed days (none)	48.1	43.6
Time since last dental visit (2 weeks - 6 months)	37.3	37.1
Time since last doctor visit (2 weeks - 6 months)	39.2	39.2

Table 5. Self-Proxy Differences for the Random Respondent Rule[1/]

Variable	Self Respondents		Persons With Proxy Reports	Difference[2/]
	1 Person Families	2+ Person Families		
1. Two-Week Recall (Rates per 100 people per quarter)				
Bed Days	182.7	102.1	226.9	-124.8*
Work Loss Days	162.5	137.8	199.6	-61.8
Cut Down Days	256.1	219.7	289.3	-69.6
Doctor Visits (Person Section)	155.4	124.8	168.4	-43.6*
Doctor Visits (Supplements)	175.5	143.7	168.4	-24.7
Dental Visits	64.4	50.1	67.6	-17.5*
Acute Conditions	137.8	98.2	119.0	-20.8*
2. 12-Month Recall (Rates Per 100 people per year)				
Hospitalizations	16.4	15.6	15.2	+0.4
3. Two-Week Recall (Percentages with at least one:)				
Bed Day	8.0	5.3	9.9	-4.6*
Work Loss Day	7.8	5.7	7.5	-1.8*
Cut Down Day	9.3	8.5	9.5	-1.0
Doctor Visit (Person Section)	17.4	14.4	16.1	-1.7
Doctor Visit (Supplements)	19.8	15.2	17.2	-2.0
Dental Visit	6.9	6.2	8.1	-1.9*
Acute Condition	17.8	13.3	17.0	-3.7*
4. 12-Month Recall (Percentages with at least one:)				
Chronic Condition	37.4	33.7	29.6	+4.1*
Hospitalization	11.5	13.0	11.8	+1.2
Doctor Visit	76.9	75.7	73.2	+2.5
Approximate n	734	1345	1795	3140

1/Estimates adjusted for the existence of multiple telephone numbers in a household and the unequal chance of selection as the "random respondent." For random respondents,
Weight = [(number of eligible adults in family) / (number of telephone numbers)].
For persons with proxy reports,
Weight = [(number of eligible adults in family - 1) / (number of eligible adults in family)] [1/(number of telephone numbers)]. Estimates based on both the control and experimental questionnaire data.

2/Difference = (Self respondents in 2+ families) - (persons with proxy reports).

* Significant difference at .05 level, using standard errors reflecting the complexity of sample design.

Table 6: Estimated Response Rates:
NCHS RDD Telephone Surveys

	Household		Person		Respondent Rule
	N	RR	N	RR	
NTHIS - 1	1,150	83%	2,300	76%	Self-all 17+ in HH
NTHIS - 2	5,500	86	11,000	78	Self-all 17+ in HH
NTHIS - Longitudinal					
- Wave 1	4,301	84	8,999	73	Self-all 17+ in HH
- Wave 2	3,741	(87)	7,963	(89)	Wave 1 respondents
Health Practices - Longitudinal					
- Wave 1	-	-	3,025	81	Random adult for self
- Wave 2	-	-	2,436	(81)	Wave 1 respondents
SRC - Knowledgeable adult	2,340	82	-	-	Knowlegeable adult for all HH adults
- Random adult	2,134	78	-	-	Random adult for all HH adults
Johns Hopkins (Baltimore SMSA)	2,474	76	-	-	Knowledgeable adult for all persons in HH

FIGURE 1

O.M.B. No. 0937-0021: Approval Expires June 30, 1983

FORM HIS-1T(1982)
(5-6-82)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR
U.S. PUBLIC HEALTH SERVICE

NATIONAL HEALTH INTERVIEW SURVEY

TELEPHONE AVAILABILITY AND STATUS SUPPLEMENT

NOTICE – All information which would permit identification of the individual will be held in strict confidence, will be used only by persons engaged in and for the purposes of the survey, and will not be disclosed or released to others for any purposes.

1. Book _____ of _____ Books

2. R.O. number [3,4]

3. Sample [5–7]

5. Control number — PSU [8–10] | Segment [11–14] | Serial [15–16]

13. Interviewer's name — Code [17,18]

11a. Is there a telephone in this [house/apartment]? [19]
- 1 ☐ Yes
- 2 ☐ No (11g)

b. What is the telephone number? [20–29]
Area code/Number
_____ / _____

c. Is this number for business only? [30]
- 1 ☐ Yes
- 2 ☐ No

d. Are there any other telephone NUMBERS in this [house/apartment]? Do not include telephone extensions. [31]
- 1 ☐ Yes
- 2 ☐ No (12)

e. What are the other telephone numbers in this [house/apartment]?
Enter EACH number including area code.
Area code/Number

		f. Is (number in 11e) for business only?		
		Yes	No	
_____ / _____	[32–41]	1 ☐	2 ☐	[42]
_____ / _____ (111)	[43–52]	1 ☐	2 ☐	[53]
_____ / _____	[54–63]	1 ☐	2 ☐	[64]

ASK for EACH number in 11e.

After asking 111 for each number, go to Household page p.12.

g. Is there a telephone elsewhere on which people in this [house/apartment] can be called when they *are* at home? [65]
- 1 ☐ Yes
- 2 ☐ No (12)

h. What is the telephone number? [66–75]
Area code/Number
_____ / _____

i. Where is that telephone located? [76–77]
Be specific.
Location
_____
_____ (12)

FOOTNOTES

Indirect Standardization of Neonatal Mortality for Birth Weight

Joel C. Kleinman
Division of Analysis
National Center for Health Statistics

INTRODUCTION

Recent advances in perinatal medicine have led to an unprecedented decline in mortality among low birth weight infants.[1,2] These advances, together with the proliferation of regionalized perinatal networks, have increased the need for surveillance and monitoring of trends and variations in birth weight-specific perinatal mortality rates. Ideally, this type of monitoring requires a national system of linked birth and death certificates. Although several States do have such systems available, there is no national data base in the United States which can be used to develop estimates of weight-specific mortality rates.

The indirect method of standardization provides a measure (the Standardized Mortality Ratio, SMR) which may be useful for summarizing a set of weight-specific neonatal mortality rates. The SMR can be applied without a national system of linked birth-death records since it requires only the total number of deaths and the distribution of births by weight. Such a measure has been used, in fact, to compare perinatal outcomes among administrative areas in England and Wales[3-7] and California hospitals and counties.[8-10] The purpose of this presentation is to summarize previously reported results concerning the properties of the indirect method empirically using geographic variation among 202 Health Service Areas in the United States. Three issues are discussed: (1) the usefulness of the SMR when the weight-specific mortality rates are not a constant multiple of the standard rates, (2) estimation of the standard error of the SMR, and (3) examples of how the SMR varies geographically and among hospitals.

METHODS

Neonatal mortality (for a hospital or geographic area) can be viewed as the outcome of two separable processes: the distribution of birth weight in that population and the probabilities of death within each birth weight group. In particular, let

B_i = number of births in weight group i during the year

$B = \Sigma B_i$

$p_i = B_i/B$

D_i = number of deaths among births in weight group i

$D = \Sigma D_i$

$\hat{m}_i = D_i/B_i$

Then,

$$\hat{m} = \frac{D}{B} = \frac{\Sigma B_i \hat{m}_i}{B} = \Sigma p_i \hat{m}_i = (\Sigma p_i M_i)\left(\frac{\Sigma p_i \hat{m}_i}{\Sigma p_i M_i}\right) = \hat{M}\,\hat{\theta} \qquad (1)$$

where the M_i represent a set of standard weight-specific mortality rates. The last term on the right, $\hat{\theta}$, is the standardized mortality ratio (SMR) and the next-to-last term, $\hat{M}$, is the expected neonatal mortality rate (ENMR). The ENMR is a summary measure which can be used to compare the birth weight distributions of different populations. Similarly, the SMR can be used to compare the weight-specific mortality rates experienced by different populations, holding their birth weight distributions constant.

The SMR is most useful when the area's weight-specific mortality is a constant multiple of the standard for all birth weights, $m_i/M_i = \theta$. In this situation the ratio of SMR's for two areas is a complete summary of the relative difference in survival rates between the two areas. However, when m_i/M_i is not constant the SMR has two important limitations. First, consider the case of two areas which have the same set of weight-specific rates, but these rates are not a constant multiple of the standard. If each of these areas has a different birth weight distribution then their SMR's will not be equal. An example of the second limitation is the situation in which one area has lower mortality among certain birth weights but higher mortality among other weights. The use of a single summary measure like the SMR cannot reflect this variation.

If linked birth and death records are available, analysis of SMR's should be considered only a first step. Intensive examination of birth weight-specific mortality rates should be used to pinpoint specific problems. For example, if an area's SMR is high, it is important to determine whether mortality is high at all birth weights or whether the problem is confined to, say, infants in the 1,000 to 1,500 gram category. Intervention strategies in the two cases may differ. In addition, mortality for a specific weight group could be high while the overall SMR is not much above the average.

Nevertheless when linked birth and death records are unavailable the SMR can be used to assess variations in survival using only the weight distribution of live births and the total number of deaths. Furthermore, the SMR can always be interpreted as a measure of excess (or deficit) mortality compared to the standard.

When linked birth and death certificates are unavailable, analysis of crude or standardized neonatal mortality rates requires certain assumptions. In particular, the information on the birth and death certificate must be assumed comparable in terms of race and residence. Furthermore, the neonatal deaths used in the numerator are assumed to derive from the births in the denominator. While this assumption is not strictly true (e.g., births occurring in December of one year may contribute to deaths in January of the following year), it should not introduce much error when fertility rates do not change rapidly.

Standard Error Estimates

Since the ENMR's and SMR's will be used to investigate geographic variation among rather small geographic areas, it is necessary to estimate their standard errors (SE's). The ENMR will be used to summarize the area's birth weight distribution. Thus, its SE is derived under a model which assumes that the B_i are the realization of a multinomial random process with B "trials" and underlying probability π_i of a birth falling into weight group i. The SE of the SMR is derived conditionally upon the B_i since the SMR is used to compare weight-specific mortality given the birth weight distribution. Thus the number of deaths in a specific birth weight category is assumed to be a binomial random variable with parameters B_i and m_i. The SE estimates are then obtained from straightforward (but tedious) algebra. More details are given in other articles.[11,12]

Since relative values of ENMR's and SMR's are of primary interest (i.e., two areas should be compared in terms of the ratio of their ENMR's or SMR's), these formulas are presented in terms of logarithms:

$$s(\ln\hat{M}) = \sqrt{\frac{1}{B}\left(\frac{\hat{Q}-\hat{M}}{\hat{M}}\right)}$$

$$s(\ln\hat{\theta}) = \sqrt{\frac{1}{D}\,(1-\hat{Q})} \qquad (2)$$

$$\hat{Q} = \Sigma B_i M_i^2 \Big/ \Sigma B_i M_i$$

The use of a log transformation allows the ratio of two ENMR's or SMR's to be expressed as the difference in their logarithms. For example, $\ln(\hat{\theta}_1/\hat{\theta}_2) = \ln\hat{\theta}_1 - \ln\hat{\theta}_2$. The SE of this difference is then easily calculated as:

$$\sqrt{s^2(\ln\hat{\theta}_1) + s^2(\ln\hat{\theta}_2)}$$

Confidence intervals or significance tests based on the log transformation also have the advantage of being independent of which area is chosen for the denominator of the ratio.

The estimated standard error of the SMR has a bias of about 5 percent in absolute value.[11] It is important to note that this estimate results in a lower SE than the approximation suggested by Mallett and Knox[6] who ignore the effect of the factor $1-\hat{Q}$. Ignoring the factor $1-\hat{Q}$ results in the SMR and neonatal mortality rate (NMR) having the same standard error. However, the mean value of $\sqrt{1-\hat{Q}}$ over the HSA's was .81 to .88 for whites and .74 to .80 for blacks (depending upon the standard), indicating that the SMR has a substantially smaller SE than does the NMR.

Data Sources

The National Center for Health Statistics (NCHS) assembles data from birth and death certificates for 56 reporting areas covering the entire United States. These data were aggregated by geographic area into 202 Health Service Areas (HSA's). The HSA's were defined in accordance with the Health Planning and Resources Development Act of 1974.[13] Minor modifications of the official HSA's were required to maintain county boundaries when aggregating the vital statistics. The natality data used provided birth weight coded in 10 categories (500 gram intervals). We also combined data for the years 1974-77 in order to reduce the effects of random variation resulting from small numbers of births and deaths. Data for white and black infants were always considered separately. Tables for black infants were based on the 156 HSA's with 1,000 or more black births during the 4-year period. All 202 HSA's had more than 1,000 white births.

Three sources of data were utilized to obtain weight-specific rates which could be used as the standards in computing SMR's.[11] The weight-specific rates from these sources are shown in table 1. Each set exhibits the characteristic "cross-over" between white and black rates: blacks have lower rates below about 3,000 grams and higher rates above 3,000. Taken as six separate sets of weight-specific rates they exhibit a wide range of values which we will use to investigate the properties of the SMR.

RESULTS

Variation in SMR when Rates are Equal

One of the major limitations of the SMR is that when the weight-specific mortality rates are not a constant multiple of the standard rates (m_i/M_i not constant), SMR's for two areas may not be equal even if their weight-specific mortality rates are identical. In order to empirically investigate the extent to which this occurs, the following procedure was used. Three pairs of the weight-specific rates in table 1 were applied to the actual birth weight distribution of each HSA. One set is used as the "true" rates to determine the number of neonatal deaths in the HSA and the other is used as the standard.

Three situations were considered in these simulations: (1) the 1960 U.S. rates were used as the standard and the 1973-74 North Carolina rates as the true rates; (2) the 1960 U.S. rates as standard and the 1977 California rates as the true; and (3) the 1973-74 North Carolina rates as standard and the 1977 California rates as the true.

Since the same weight-specific mortality rates are applied to each HSA, an ideal summary measure of weight-specific mortality should give the same value to all HSA's for each of the three cases studied. But since m_i/M_i is not constant the SMR's will in fact vary. The extent of this variation over the HSA's reflects the imperfection of the SMR as a summary measure of weight-specific mortality.

Since the under 500 gram group has a mortality rate of 1 we modified the SMR by omitting this group from the SMR, i.e., calculating the SMR as:

$$\hat{\theta} = \frac{D - B_1}{\Sigma M_i B_i - B_1}$$

By eliminating these infants from the SMR we obtain a more sensitive measure of preventable mortality.[11]

Table 2 shows the distribution of SMR's using the different sets of standard and "true" rates. It is apparent that the distributions for whites show much less variation than do those for blacks. In any case,

virtually all the HSA's had SMR's within 2 percent of the median for whites and within 3 percent of the median for blacks.

Distribution and Relationship Among Measures

Table 3 shows the correlation coefficients for selected measures using the two most current standard rates (NC73-4 and CA77) for each race. For whites, the ENMR's have a slight negative correlation with the SMR's. There is also a substantial correlation between the NMR and the SMR's. In terms of the identification of areas with excess mortality, the NMR and SMR produce different results despite the high correlation. For example of the HSA's among the highest quintile for white SMR's (NC73-4 standard), 33 percent were not in the highest quintile for the NMR. For blacks, 41 percent of the HSA's in the highest quintile based on the SMR were not in the highest quintile based on the NMR.

It is also of interest to note that, for both black and white infants, the proportion of births below 1,500 grams (VLBW) is highly correlated with the ENMR's but the proportion below 2,500 grams (LBW) is not. The correlation between VLBW and LBW is also modest: .47 for whites and .69 for blacks. This supports other suggestions that the proportion of births below 1,500 is a better indicator of expected mortality than the usual low birth weight cutoff of 2,500 grams.[14]

Table 4 shows the percentile distributions of the ENMR, SMR (using NC73-4 standards), NMR and VLBW rate. There is substantial variation in all the measures, although the greater variation among black measures is due in part to their larger standard errors (i.e., smaller number of births upon which measures are based). There is a 51 percent spread between the 5th and 95th percentiles for neonatal mortality among whites and a 97 percent spread among blacks. The corresponding spreads for SMR's are somewhat smaller (43 percent among whites and 87 percent among blacks) but those for ENMR's considerably smaller (30 percent among whites and 46 percent among blacks).

These data, together with the correlations in table 4, suggest that geographic variation in neonatal mortality among HSA's is due in larger part to variations in weight-specific mortality than to variations in birth weight distribution. When ln(SMR) and ln(ENMR) are entered into a multiple regression on ln(NMR), the standardized regression coefficient for ln(SMR) is greater than that for ln(ENMR) (.90 versus .67 among whites, .75 versus .65 among blacks, using NC73-4 standards).

Use of SMR's to Evaluate Hospital Performance

The SMR can also be very useful in evaluating the effectiveness of perinatal care.[8,15] For example, in a study[15] of variations in SMR's among eight District of Columbia hospitals we used a generalized least squares approach to model the trend in the logarithm of SMR's over three time periods from 1970-78. We started with a saturated model for the 8 X 3 table of SMR's (table 5). Model parameters representing hospital-specific changes were sequentially deleted (or combined) if they were not significant at the 10 percent level.

Table 6 outlines the major steps in the model fitting procedure. There were no significant changes in SMR's between 1970-72 and 1973-75 (lines 1b and 2a). However, the changes in SMR during the 1973-75 to 1976-78 time period did differ significantly among the hospitals (line 1c). The 8 hospitals could be divided into two groups which experienced similar changes in SMR's during this period (lines 3a, 3b). One group (Hospitals A, C, D, F) showed no significant change (line 4b) while the other group (B, E, G, H) experienced a reduction in SMR's of 37 percent (95 percent confidence interval: 27 percent to 46 percent).

How important is this variation in explaining the District's lack of improvement in neonatal mortality? One way to approach this question is to estimate what the District's SMR would have been if the four hospitals which experienced no change in neonatal SMR from 1973-75 to 1976-78 actually experienced a decrease equal to that of the remaining hospitals. This may be somewhat unrealistic since hospital F had the lowest SMR's in the first two periods. Thus we estimate the effects of reductions in hospitals A, C, and D only. If these three hospitals experienced the 37 percent decrease, the District's overall SMR in 1976-78 would have been reduced by 20 percent.

DISCUSSION

The previous results suggest that the SMR is a useful summary measure of weight-specific mortality rates even when the weight-specific rates are not a constant multiple of the standard rates. The exclusion of births under 500 grams further improves the sensitivity of the SMR as a measure of preventable mortality. In addition the choice of standard rates has little effect on the SMR.

Certain factors which could be used to further adjust the NMR were excluded for the following reasons:

(1) Sex and plurality. It is well known that females or multiple births are generally lighter and have different weight-specific mortality than males or single births. However, there was no reason to expect a great deal of variation among HSA's in the sex ratio or plurality within birth weight categories. Thus, the addition of these factors to the standardization procedure would have added considerably to the data processing requirements, with little change in the results. When analyzing data for smaller units such as hospitals, it may be worthwhile to use these factors in the computation of the ENMR.

(2) Gestation. On the clinical level, there are important differences between premature and small-for-date babies. Unfortunately the completeness and quality of the gestation information on birth certificates is questionable.[16] For example, 20 percent of the birth certificates do not report gestation. In fact, it has been shown that birth weight is a much better predictor of mortality than gestation and the latter adds very little improvement in prediction.[17]

Another omission of this approach relates to fetal deaths. It would certainly be preferable to use an adjusted perinatal mortality rate (or separate fetal and neonatal rates) for monitoring. Fetal deaths were not considered, however, for two primary reasons. First, there is some concern that fetal deaths are not completely or consistently reported across the U.S. Second, the quality and completeness of information on the fetal death certificate is limited. For example, in 1977, 17 percent of white and 12 percent of black fetal deaths had birth weight not stated compared to 0.2 percent for both black and white live births.

Throughout this presentation white and black births have been analyzed separately. Marshall and Mallett[18] also discuss the growing importance of racial mix on geographic variations in perinatal mortality in England. On the other hand, Williams[8-10] calculates a SMR which combines all races (adjusting for their different weight-specific mortality). It is important, however, to analyze mortality separately for whites and blacks since their birth weight distributions and weight-specific mortality rates are so different and their patterns of medical care may also be different. In fact, the correlation between white and black SMR's (based on the 156 HSA's with more than 1,000 black births) was only .54. Thus areas with high black SMR's could be missed if race-specific rates are not used.

There are several assumptions underlying the SMR which require that it be used with caution. While controlling for birth weight does control for a large portion of the variation in the risk level of the population, it is important to note that there are several variables which are not available on the birth certificate but could affect weight-specific mortality. For example, women who live at high altitudes have lighter babies.[19] The effect of this phenomenon on the SMR is illustrated clearly by white births in the Denver, Colorado HSA. This HSA had one of the highest ENMR's and one of the lowest SMR's among all HSA's. The explanation is that Denver's high altitude results in women having babies whose low birth weights are not indicative of subsequent health problems.

Despite these limitations, however, the SMR and ENMR (or VLBW rate) provide useful summary measures for monitoring trends and variations among geographic areas or medical care facilities in birth outcomes. If linked birth and death records are available, monitoring based on the SMR should be supplemented by analysis of weight-specific mortality rates.

REFERENCES

1. Kleinman JC, Kovar MG, Feldman JJ, Young CA: A comparison of 1960 and 1973 early neonatal mortality in selected states. Am. J. Epidemiol. 108:454-469, 1978.

2. Lee KS, Paneth N, Gartner LM, Pearlman MA, Gruss L: Neonatal mortality: An analysis of the recent improvement in the United States. Am. J. Public Health 70:15, 1980.

3. Brimblecombe FSW, Ashford JR, Fryer JG: Significance of low birth weight in perinatal mortality: A study of variations within England and Wales. Brit. J. Prev. Soc. Med. 22: 27-35, 1968.

4. Ashford JR, Read KLQ, Riley VC: An analysis of variations in perinatal mortality amongst local authorities in England and Wales. Int. J. Epidemiol. 2:31-46, 1973.

5. Chalmers I, Newcombe R, West R, Campbell H, Weatherall J, Lambert P, Adelstein A: Adjusted perinatal mortality rates in administrative areas of England and Wales. Health Trends 10:24-29, 1978.

6. Mallett R, Knox EG: Standardized perinatal mortality ratios: Technique, utility and interpretation. Community Medicine 1:6-13, 1979.

7. Macfarlane A, Chalmers I, Adelstein AM: The role of standardization in the interpretation of perinatal mortality rates. Health Trends 12:45-50, 1980.

8. Williams RL: Measuring the effectiveness of perinatal medical care. Med. Care 17(2):95-110, Feb. 1979.

9. Williams RL, Cunningham GC, Norris FD, Tashiro M: Monitoring perinatal mortality rates: California, 1970 to 1976. Am. J. Obstet. Gynecol. 136:559-568, Mar. 1980.

10. Williams RL: Measuring the Effectiveness of Perinatal Medical Care. University of California. Grant No. MC-R-060390-01. Prepared for the Bureau of Community Health Services, Health Services Administration. Rockville, Md., July 31, 1977.

11. Kleinman JC: Indirect Standardization of Neonatal Mortality for Birth Weight. Int. J. Epidemiol., June 1982.

12. Foster JE, Kleinman JC: Adjusting Neonatal Mortality Rates for Birth Weight. To be published in Vital and Health Statistics. Series 2.

13. Bureau of Health Manpower: The Area Resource File: A Manpower Planning and Research Tool. DHEW Pub. No. (HRA) 80-4. Health Resources Administration. Washington. U.S. Government Printing Office, Aug. 1979.

14. Less KS, Paneth N, Gartner LM, Pearlman M: The very low-birth-weight rate: Principal predictor of neonatal mortality in industrialized populations. J. Pediatr. 97(5):759-764.

15. Madans JH, Kleinman JC, Machlin SR: Differences among hospitals as a source of excess neonatal mortality: the District of Columbia, 1970-78. J. Comm. Health 7(2):103-117, Winter 1981.

16. David RJ: The quality and completeness of birthweight and gestational age data in computerized birth files. Am. J. Public Health 70:964-73, 1980.

17. Susser M, Morolla FA, Fleiss J: Birth weight, fetal age and perinatal mortality. Am. J. Epidemiol. 96:197-204, 1972.

18. Marshall T, Mallett R: The influence of racial mix on comparisons of perinatal mortality rate between Area Health Authorities. Int. J. Epidemiol. 9:255-263, 1980.

19. Meyer MB: Effects of maternal smoking and altitude on birth weight and gestation. In Reed DM and Stanley FJ (eds), The Epidemiology of Prematurity. Urban and Schwarzenberg, Baltimore, 1977.

Table 1. Weight-specific neonatal mortality rates per 100,000 live births

Birth weight	US60		NC73-4		CA77	
	White	All other	White	Black	White	Black
500 grams or less	100,000	100,000	100,000	100,000	100,000	100,000
501-1,000	89,663	86,962	89,746	86,804	68,680	61,011
1,001-1,500	57,600	44,035	46,466	32,346	24,733	13,855
1,501-2,000	18,032	10,995	15,183	9,063	6,412	4,011
2,001-2,500	3,946	2,807	3,191	1,708	1,754	961
2,501-3,000	1,102	1,110	595	510	521	434
3,001-3,500	457	671	324	343	146	215
3,501-4,000	306	609	253	313	146	174
4,001-4,500	356	836	247	586	113	179
Above 4,500	857	1,864	517	2,334	186	512

Table 2. Percent distributions of Standardized Mortality Ratio applying the same weight-specific mortality rates to each Health Service Area[a]

Rates used as Standard - "True"	White		Black	
	Including infants 500 grams or less	Excluding infants 500 grams or less	Including infants 500 grams or less	Excluding infants 500 grams or less
US60 - NC73-4				
4 percent below median	-	-	6.4	4.5
3 percent below median	0.5	0.5	7.1	3.8
2 percent below median	4.5	0.5	10.9	14.1
1 percent above or below median	92.1 (.81)	97.0 (.80)	62.2 (.77)	66.0 (.75)
2 percent above median	2.0	1.5	8.3	9.0
3 percent above median	0.5	0.5	2.6	1.9
4 percent above median	0.5	-	2.6	0.6
US60 - CA77				
4 percent below median	1.0	0.5	9.6	3.2
3 percent below median	1.5	0.5	10.9	3.2
2 percent below median	10.4	10.4	14.7	7.7
1 percent above or below median	72.8 (.50)	86.6 (.48)	46.2 (.49)	62.2 (.44)
2 percent above median	9.4	1.5	5.8	14.1
3 percent above median	3.0	0.5	6.4	6.4
4 percent above median	2.0	-	6.4	3.2
NC73-4 - CA77				
4 percent below median	1.0	0.5	4.5	1.3
3 percent below median	3.0	-	3.8	1.9
2 percent below median	15.8	0.5	15.4	5.8
1 percent above or below median	74.8 (.62)	96.5 (.59)	55.1 (.63)	82.0 (.59)
2 percent above median	4.0	2.0	8.3	7.7
3 percent above median	0.5	0.5	7.1	1.3
4 percent above median	1.0	-	5.8	-

[a]Numbers in parentheses indicate median value of SMR.

Table 3. Means, standard deviations, and correlations for selected measures[a]: U.S. Health Service Areas, 1974-77

	Mean	SD	Correlation Coefficients						
			1	2	3	4	5	6	7
White									
1. ENMR (NC73-4)	-4.393	.083	1.00	.96	-.21	-.26	.94	.68	.51
2. ENMR (CA77)	-4.870	.096		1.00	-.22	-.32	.93	.58	.45
3. SMR (NC73-4)	-.223	.121			1.00	.99	-.22	-.01	.73
4. SMR (CA77)	.294	.124				1.00	-.27	-.01	.70
5. VLBW	-4.705	.123					1.00	.47	.45
6. LBW	-2.794	.107						1.00	.44
7. NMR	-4.599	.124							1.00
Black									
1. ENMR (NC73-4)	-3.872	.139	1.00	.99	.01	-.04	.94	.70	.66
2. ENMR (CA77)	-4.326	.168		1.00	.10	-.05	.89	.63	.66
3. SMR (NC73-4)	-.240	.189			1.00	.99	.04	.12	.75
4. SMR (CA77)	.279	.188				1.00	-.07	.11	.71
5. VLBW	-3.787	.175					1.00	.69	.58
6. LBW	-2.068	.113						1.00	.53
7. NMR	-4.083	.222							1.00

[a]Log scale.

Table 4. Percentile distributions of selected measures: U.S. Health Service Areas 1974-77

Percentile	White				Black			
	VLBW	NMR	ENMR	SMR	VLBW	NMR	ENMR	SMR
	-Rate per 1,000-				-Rate per 1,000-			
5	7.3	8.1	10.7	.672	16.5	11.5	17.1	.553
10	7.8	8.5	11.1	.700	18.6	12.3	17.7	.624
20	8.3	9.1	11.6	.726	19.7	14.6	18.6	.697
30	8.6	9.5	11.9	.755	21.3	15.7	19.5	.728
40	8.8	9.9	12.2	.769	22.2	16.5	20.4	.768
50	9.1	10.1	12.4	.795	23.2	17.4	21.2	.796
60	9.4	10.4	12.6	.815	24.0	18.2	21.7	.816
70	9.6	10.6	12.9	.847	24.9	18.8	22.2	.843
80	9.9	11.0	13.1	.878	25.8	19.9	23.0	.900
90	10.3	11.6	13.5	.927	27.3	21.3	24.1	.975
95	10.7	12.2	14.0	.961	28.9	22.7	24.9	1.036

Table 5. Crude neonatal mortality rate per 1,000 live births, and observed and fitted Standardized Mortality Ratios by hospital and time period: D.C. nonwhite resident births and deaths

	Crude Neonatal Mortality			Observed SMR			Fitted SMR[a]		
	1970-72	1973-75	1976-78	1970-72	1973-75	1976-78	1970-72	1973-75	1976-78
All nonwhite resident births	20.3	20.3	19.8	100	110	92	-	-	-
Hospital A	28.7	23.7[b]	33.5	121	115	111	114	114	114
Hospital B	17.2	15.1	12.3	103	102	62	102	102	64
Hospital C	23.6	28.8	30.2	118	120	126	119	119	119
Hospital D	20.8	19.9	22.3	111	107	113	112	112	112
Hospital E	13.5[b]	23.1[b]	16.1[b]	81[b]	116[b]	76[b]	101	101	64
Hospital F	8.9[b]	11.9[b]	16.7[b]	65[b]	77[b]	96[b]	83	83	83
Hospital G	14.3[b]	18.8[b]	7.7[b]	88[b]	116[b]	53[b]	98	98	62
Hospital H	21.7	19.9	17.5	127	133	86	133	133	84

[a]Fitted model based on stepwise generalized least squares procedure. Goodness of fit: $X^2 = 9.23$ with 15 df. See text for details.

[b]Coefficient of variation greater than 15 percent.

Table 6. Outline of stepwise model fitting procedure

Hypothesis test given specified model (Model)	df	X^2	p
1. Saturated model			
a. all $H_iT_1=H_iT_2=0$	14	26.47	.023
b. all $H_iT_1=0$	7	5.12	.645
c. all $H_iT_2=0$	7	25.71	.001
2. Model with all $H_iT_1=0$*	7	5.12	.645
a. $T_1=0$	1	.53	.468
3. Model with all $H_iT_1=0$, $T_1=0$*	8	5.65	.687
a. Grp 1: $H_AT_2=H_CT_2=H_DT_2=H_FT_2$	3	3.72	.293
a. Grp 2: $H_BT_2=H_ET_2=H_GT_2=H_HT_2$	3	2.42	.491
4. Model with all $H_iT_1=0$, $T_1=0$, Grp 1, Grp 2*	14	11.10	.678
a. Grp 1 = Grp 2	1	34.90	.000
b. Grp 1 = 0	1	1.04	.308
c. Grp 2 = 0	1	43.38	.000
5. Final model: all $H_iT_1=0$, $T_1=0$, Grp 1 = 0*	15	12.14	.669

*These lines give the goodness-of-fit X^2 for the specified model.

Key

Symbol	Description
T_1 =	Change between the time periods 1970-72 and 1973-75
T_2 =	Change between the time periods 1973-75 and 1976-78
H_i =	Main effect for hospital i (i = A-H)
H_iT_j =	Interaction between i^{th} hospital and j^{th} time interval
Grp 1 =	H_iT_2 interactions for hospitals A, C, D, and F
Grp 2 =	H_iT_2 interactions for hospitals B, E, G, and H

Monthly Patterns of Marital Relations and Waiting Times to Conception

Gerry E. Hendershot
National Center for Health Statistics

INTRODUCTION

Women who attempt to conceive by scheduling intercourse for the most fecundable part of the month are older and less fecund than noncontracepting newlyweds. Also, they are slower to conceive, indicating that their scheduling of intercourse only partly compensates for their lower fecundity. These are the main findings of the analysis presented here, which is based on data from the National Survey of Family Growth, conducted by the National Center for Health Statistics in 1976.

The statistical study of waiting times to conception began more than half a century ago with the work of Corrado Gini.[1] Popular interest in the subject has revived because more women are postponing childbearing to ages at which fecundity, the ability to conceive, is lower. In a recent French study it was reported that 73-74 percent of women under age 31 years conceived within a year, but that figure dropped to 61 percent for women 31-35 years of age, and 54 percent for women over 35.[2] The women in the study were married to men who were completely sterile and they were artificially inseminated, so the decline in fecundity could not be attributed to the infecundity of their husbands nor to a decline in the frequency of their intercourse. Although critics of the study have questioned its reports of low levels of fecundity and rapid decline at ages 30-34, they agree with its conclusion that fecundity does decline after age 30.[3]

The number of U.S. women whose childbearing plans are affected by that decline in fecundity is increasing. The percent of wives 30-34 years of age who expected to bear children increased from 14 percent in 1971[4] to 17 percent in 1979[5] and the number of such women increased from 0.5 million to 0.9 million in the same period. The increase was even more dramatic for childless white wives aged 30-34, from 24 percent or 50,000 expecting children in 1971 to 44 percent or 275,000 in 1979. Thus, in only eight years the number of would-be mothers passing the age at which fecundity decline accelerates, increased more than fivefold. These increases probably resulted from the growing labor force participation of women: the percent of women who worked between marriage and the birth of their first child increased from 66 percent among women married in 1955-59 to 80 percent among women married in 1965-69;[6] and women who work tend to postpone the first birth for a longer interval than other women.[7]

So many women have postponed childbearing so long that they must now achieve higher first- and second-order birth rates at ages 30-34 years than nearly all previous U.S. cohorts if they are to have all the children they expect.[8] To achieve such high fertility despite their declining fecundity, many women must take special measures to conceive. The measure most often recommended is to increase the frequency of intercourse in the period just before ovulation, which normally occurs about two weeks after menstruation begins. According to the author of a manual on infertility, "although a rigid schedule for intercourse is unrealistic and anxiety provoking, at least the frequency and approximate timing of intercourse may need to be guided somewhat by an understanding of when the sexual act is more or less likely to lead to pregnancy."[9] In theory, accurately scheduled intercourse should shorten the waiting time to conception considerably, but its actual effect, as it is practiced in the U.S. population, has not been estimated. The principal purpose of this paper is to make such an estimate.

DATA AND METHODS

The data for this study are from the National Survey of Family Growth, Cycle II (NSFG). The NSFG is a periodic survey of U.S. women 15-44 years of age. It is based on an area probability sample of the household population in the 48 conterminous states. One hour interviews are conducted in the homes of sample women by professional interviewers using a standard questionnaire. Information is obtained about marriages, pregnancies, contraception, fecundity, birth expectations, reproductive health services, and socioeconomic characteristics. The fieldwork for Cycle II of the NSFG was conducted by Westat, a commercial survey organization, under a contract with the National Center for Health Statistics. During January through September of 1976 a total of 8,611 women were interviewed, about 89 percent of those who had been identified as eligible. Women were eligible if they were 15-44 years of age and either currently married, formerly married, or never married with offspring in the household. To estimate statistics for the population represented by the sample, weights were computed for each case, reflecting its probability of selection, the rate of nonresponse for similar women, and a post-stratification adjustment to population subtotals derived independently by the Current Population Survey. Details of the sampling design and estimating procedures have been described in another publication.[10]

The principal purpose of the analyses reported here was to estimate the waiting time to conception among couples who scheduled intercourse as a means for hastening conception. Women who scheduled intercourse were identified by the following question:

> In the past three years, have you used calendar or temperature methods of rhythm to increase your chances of becoming pregnant? That is, in order to know the days when you are most likely to become pregnant?

Women who answered "yes" were asked to identify the starting and stopping months for intervals of scheduled intercourse. Those intervals were used to estimate waiting times to conception by a life table method described in the appendix. There were 321 women in the sample who reported they had used rhythm to increase the chances of pregnancy; 17 of these were missing some of the other information needed to estimate waiting times, leaving 304 usable

cases. Those 304 sample women represented about 1.4 million women in the population.

It is important to note that classifying a couple as having scheduled intercourse to conceive does not imply, necessarily, that they were following the theoretically optimum schedule. Some couples are misinformed about the most fecundable period of the month; Silber[9] notes, for instance, that some couples time intercourse to coincide with ovulation, whereas it should occur shortly before ovulation to allow time for the sperm to reach the fallopian tubes. Even if women are well informed, they may have irregular cycles, making it very difficult to predict ovulation with any precision. Or, a couple's work or travel may interfere with their scheduling of intercourse for the most fecundable time, even when it is known. Thus, "using rhythm to increase your chances of becoming pregnant" may include some practices which are less than optimal for achieving their purpose. While a medically supervised clinic population might use scheduled intercourse more effectively, this study evaluates the effectiveness of the technique as it is actually practiced in the general population.

In order to evaluate the effectiveness of scheduled intercourse as practiced by these couples, it was desirable to compare their waiting times to conception with those of a standard, "normally" fecund population. A standard population should have several characteristics: it should not be self-selected for unusually high or low fecundability; it should not be practicing contraception; it should have a relatively high frequency of intercourse; and it should not be scheduling intercourse, at least not with reference to the probability of conception. To approximate that standard population couples were selected who first married in 1971-75, were still married at the time of interview (1976), and had not used rhythm to enhance conception in the last three years. For those couples intervals were selected which were during the three years before interview and in which the couple had been fecund and having intercourse, but had not been contracepting. Because these couples were recently married (about two and one-half years on average) it is unlikely that they would have had enough reproductive experience to have formed an opinion about their own fecundity; therefore, they are not likely to have been self-selected into the population of noncontraceptors on the basis of their low fecundity, the most likely kind of bias in a study such as this. Also, because they are in the first years of a first, and so far successful, marriage, they are likely to have relatively high rates of intercourse, lessening the likelihood that that factor could account for differences in waiting times. These couples did not schedule intercourse to conceive, and in the intervals selected for study they were fecund and having intercourse, but not contracepting, so they presumably had a "normal" chance of conception. There were 450 sample women who satisfied the conditions for the standard population. Of these, 28 were missing data required for the life table analysis, leaving 422 usable cases, representing a population of about 1.7 million women.

Table 1 shows the two populations, those who scheduled intercourse and the newlywed noncontraceptors, distributed by selected characteristics: age, education, occupation, fecundity status (at interview), and use of medical services for infertility in the last three years. The two populations differ significantly (at the 5 percent level using a two tailed t-test) on all five characteristics. Women who scheduled intercourse to conceive were older, better educated, and had higher-level occupations than the newlywed noncontraceptors. They were also two and one-half times more likely to report some impairment to their fecundity (or their husband's) at the time of interview, an impairment which may well have been discovered as a result of their attempt to conceive. (Impaired fecundity includes involuntary sterility or difficulty conceiving or carrying a pregnancy to term; See Mosher, 1980.[11]) Finally, one half of the women who scheduled intercourse to conceive had talked to a professional about an infertility problem in the last three years, a figure five times higher than that for newlywed noncontraceptors, and about seven times higher than that for all nonsterile couples.[12]

Clearly, couples who schedule intercourse to conceive are a very select population, especially with respect to their fecundity. While not all of the wives, nor perhaps even a majority, have postponed childbearing to pursue a career, that type of woman seems to be disproportionately represented. The prevalence of fecundity impairments among couples in this population, and their frequent use of infertility services, both of which are probably related to their age, has two important implications for this study: first, as a group they are likely to be well informed and highly motivated users of the rhythm method to conceive; and second, the obstacles to their conception, in the form of low fecundability, are severe. It is not possible to predict, a priori, whether their presumably careful practice of rhythm will be sufficient to overcome their fecundity impairments, yielding waiting times comparable to the "normal" population of newlywed noncontraceptors.

FINDINGS

Table 2 shows estimates made by a life table technique (see appendix) of the cumulative probabilities of conception (expressed as a percent) after 2, 4, 6, 8, 10, and 12 months, for newlywed noncontraceptors and women who scheduled intercourse for the most fecundable part of the month. Also shown are the numbers of sample cases and population estimates on which the conception rates were based; these are the numbers of women who began the month of exposure identified in the column heading. Standard errors for each of the cumulative conception rates, estimated with the "jackknife" technique[13] are also shown.

Among couples who schedule intercourse for the most fecundable part of the month, 61 percent conceive in 12 months. Among newlywed noncontraceptors, however, 76 percent conceive within a year. Furthermore, at every duration of exposure up to 12 months, those who schedule intercourse are significantly less likely to have conceived. Although they have considerable success in conceiving despite their handicaps of greater age and lower fecundity, rhythm users fall significantly short of the performance of a "normal" population of younger, more fecund couples.

But are the newlywed noncontraceptors used as a standard in this study really a "normally" fecund population? The most comparable estimates of the proportions conceiving in 12 months are those by Westoff and Ryder[14] from the 1970 National Fertility Study. For married, low-parity women in their 20's, they estimated that 43-56 percent would conceive in 12 months of unprotected intercourse. That is much below the estimate in table 2 of 76 percent for newlywed noncontraceptors, and even below the estimate of 61 percent for women using rhythm to conceive. The same authors also estimated fecundability from the 1965 National Fertility Study to be much lower than estimated here.[15] At the other extreme, Silber[9] shows estimates (attributed to Westoff) which suggest much higher fecundability: in "normal" young populations between 86 and 97 percent of women conceive in less than 12 months of noncontraception. Depending on which of these several estimates reflects "normal" population fecundity, conclusions about the effectiveness of scheduled intercourse to achieve conception could differ greatly.

Fortunately, the differences may not be so great as they seem, or at least they may result from differences in assumptions and techniques rather than underlying facts. In the case of the Westoff and Ryder estimates for 1970, for instance, the authors themselves believe they are too low. Without being able to prove it, they surmise that women in the 1970 National Fertility Study did not report some of their conceptions; at a time when abortions were illegal in most parts of the nation, women may have been reluctant to mention a pregnancy whose termination was induced. Furthermore, Westoff and Ryder suggested that some women who were included in the analyses as noncontraceptors may actually have been contracepting; either they were using some method they did not think of as contraceptive, such as douching soon after intercourse, or they purposely concealed use of a method considered illicit by their religion. If they concealed pregnancies and practiced contraception, their observed conception rate would be an underestimate, as would the percent who had conceived in 12 months. The much higher percents estimated from the 1976 NSFG data may simply reflect a substantial improvement in reporting of pregnancies ending in induced abortion, and of the use of contraception.

Although Silber does not explain the derivation of the very high estimates of the percent conceiving within a year which he quotes, they almost certainly were made assuming the population was homogeneous with respect to fecundity; that is, it was assumed that all members of the population in a certain age range had the same monthly probability of conception. In a homogeneous population with a normal monthly probability of conception for young people, say 20 percent per month, the percent conceiving within 12 months is simply $1-(1-0.2)^{12}$, or 93 percent, the estimate given by Silber for those conditions. But according to Leridon,[16] populations are not homogeneous with respect to fecundability; rather, the density function of fecundability resembles a beta distribution. While many individuals in a population have fecundability close to the population average, there is a substantial minority whose fecundability is much lower. Most of those with average or above average fecundability conceive within the first few months, but those with below average fecundability tend to linger on without conceiving; consequently, the percent conceiving after a year will be considerably lower in an actual population than in a hypothetical, homogeneous population, even when their average fecundability is the same. The estimates given by Silber are correct for a homogeneous population, but for an actual population like that represented by the NSFG, Silber overestimates the percent who conceive in a year.

If the estimates of Westoff and Ryder are too low, and those of Silber are too high, those presented here may be near the mark. Confirmatory evidence is provided by the French study mentioned earlier.[2] The subjects of that study were married to completely sterile husbands, so there is no reason to suspect that they themselves were self-selected with respect to their own fecundability; they should have been, therefore, reasonably representative of the population with respect to fecundity. The pattern of "intercourse" was closely controlled, because the subjects were artificially inseminated by medical teams at the most fecundable times in their monthly cycles. Observation of the outcome was very good, of course in this clinical setting. In many ways the French study situation is like that of a normally fecund population of women whose frequency of intercourse is high; that is, the situation was like that of the newlywed noncontraceptors in this study. It is encouraging, therefore, that the estimated percent of women conceiving in a year are similar in the two studies: in the French study, 73-74 percent for women in their 20's; in this study, 76 percent for newlywed noncontraceptors, nearly all of whom are in their 20's.

If the experience of newlywed noncontraceptors can be accepted as a standard for "normal" population fecundity, the finding of this study is that the population of women who schedule intercourse to increase the probability of conception have somewhat lower fecundity. This does not mean, of course, that scheduling intercourse to conceive has no effect. On the contrary, given the advanced age and frequency of reported fecundity impairments in the population of rhythm users, their conception rate, 61 percent in the first year, is higher than might be expected. That their rate is lower than that of a "normally" fecund population indicates that their efforts to increase the probability of conception do not offset completely their fecundity handicaps.

APPENDIX

The life table technique used to estimate the waiting time to conception was based on the work of Potter.[17] The first step, which was already described, was to select intervals of scheduled intercourse or noncontraception for analysis. Then the duration of each interval was computed. The beginning of an interval was the date women began to schedule intercourse or to have unprotected intercourse; its ending came either at conception, 2 months before the interview (to avoid undetected conceptions), or when stopped for some other reason, such as beginning to use contraception. The intervals then were classified by their duration in months.

The next step was to estimate the number of woman-months of exposure to the risk of pregnancy which occurred at each month of duration. It was assumed that all women who completed a month of observation without conceiving or otherwise dropping out each contributed one full month of exposure in that month. Also, those who conceived in that month were counted as each contributing a full month of exposure. But those who dropped out for other reasons, either through termination of observation (due to interview) or any other reason, were counted as having contributed only one-half month of exposure each. If for example there were 100 intervals of six months duration or longer, 20 intervals which ended in pregnancy during the sixth month, and 10 more which ended for other reasons, then the total months of exposure during the sixth month was 100 + 20 + (10/2) = 125.

The monthly probability of conception, or fecundability, was then estimated by dividing the number of pregnancies occurring in each month of duration by the number of woman-months of exposure in that month. In the example, fecundability in the sixth month is the number of pregnancies in the sixth month, 20, divided by the woman-months of exposure in the sixth month, 125, or 0.16. The probability of not conceiving in a month is the complement of fecundability for that month; in the example, the probability of not conceiving in the sixth month is 1 - 0.16 = 0.84.

To estimate the percent of women conceiving by the end of any particular month of scheduled intercourse or noncontraception, the probability of not conceiving by that time was first estimated. The probability of not conceiving by the end of any paricular month was found by multipling together all the probabilities of not conceiving in each individual month from the first through the month in question. For example, if the probabilities of not conceiving in the first through the sixth months were, respectively, 0.70, 0.75, 0.80, 0.74, 0.83, and 0.84, then the probability of not conceiving by the end of the sixth month is the product of those probabilities, or 0.70 x 0.75 x 0.80 x 0.83 x 0.84 = 0.29. This may be interpreted as meaning that 29 percent of women exposed to these risks of conception would not have conceived in six months. By taking the complement of the probability of not conceiving after a given interval, we have the desired statistic, the probability of conceiving after a given interval. In the example, this is 1 - 0.29 = 0.71, which can be interpreted to mean that 71 percent of women would have conceived after six months of this regimen.

REFERENCES

1. Gini, C. (1924) Premieres Recherches sur La Fecondabilite de la Femme. Proc Math Congr 1924: 889-892.

2. Federation CECOS, Schwartz, D., and Mayaux, M.J. (1982) Female Fecundity as a Function of Age. The New England Journal of Medicine. 306: 404-406.

3. Bongaarts, J. (1982) Infertility after Age 30: A False Alarm. Family Planning Perspectives. 14: 75-78.

4. U.S. Bureau of the Census. (1974) Fertility Histories and Birth Expectations of American Women: June 1971. Current Population Reports. Series P-20, No. 263.

5. U.S. Bureau of the Census. (1980) Fertility of American Women: June 1979. Current Population Reports. Series P-23, No. 358.

6. Bumpass, L.L. and Sweet, J.A. (1980) Patterns of Employment Before and After Childbirth. Vital and Health Statistics. Series 23, No. 4.

7. Groat, H., Wicks, J., Neal, A.G., and Hendershot, G.E. (1982) Working Women and Childbearing. Vital and Health Statistics. Series 23, No. 9.

8. Campbell, A.A. (1981) Needed Research on Birth Expectations. Pp. 291-304 in Hendershot, G.E. and Placek, P. (eds.) , Predicting Fertility: Demographic Studies of Birth Expectations. Lexington, Massachusetts: Lexington Books.

9. Silber, S.J. (1980) How to Get Pregnant. New York, New York: Warner Books.

10. Grady, W.R. (1981) National Survey of Family Growth, Cycle II: Sample Design, Estimation Procedures, and Variance Estimation. Vital and Health Statistics. Series 2, No. 87.

11. Mosher, W.D. (1980) Reproductive Impairments Among Currently Married Couples: United States, 1976. Advance Data. No. 55.

12. Hendershot, G.E. and Bauman, K.E. (1981) Use of Services for Family Planning and Infertility: United States. Vital and Health Statistics. Series 23, No. 8.

13. McCarthy, J. (1978) A Comparison of the Probability of the Dissolution of First and Second Marriages. Demography. 15: 345-359.

14. Westoff, C.F. and Ryder, N.B. (1977) The Contraceptive Revolution. Princeton, New Jersey: Princeton University Press.

15. Ryder, N.B. and Westoff, C.F. (1971) Reproduction in the United States, 1965. Princeton, New Jersey: Princeton University Press.

16. Leridon, H. (1977) Human Fertility, the Basic Components. Tr. Helzner, J.D. Chicago, Illinois: University of Chicago Press.

17. Potter, R.G. (1966) Application of Life Table Techniques to Measurement of Contraceptive Effectiveness. Demography. 3: 297-304.

Table 1. Number of sample women, estimated number of women in the population and percent distribution by age, education, occupation, fecundity status, and use of medical services for infertility in the last 3 years, according to monthly pattern of marital relations

Age, education, occupation, fecundity status, and use of medical services for infertility in the last 3 years	Monthly pattern of marital relations	
	Scheduled intercourse	Newlywed noncontraceptors
	Sample women	
All women	304	422
	Population in thousands	
All women	1,389	1,659
	Percent distribution	
Age		
15-29 years	69	92
30-44 years	31	8
Education		
High school or less	51	74
More than high school	49	26
Occupation		
Professionals and managers	29	14
Other or none	61	86
Fecundity status		
Fecund	60	84
Impaired fecundity	40	16
Use of medical services for infertility in the last 3 years		
Did not use services	50	90
Used services	50	10

1 See text for definitions of monthly patterns of marital relations.

NOTE: The differences between percents for the monthly patterns in each line are statistically significant at the 0.05 level or below.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1982-361-166:341

HRST
From the Office of Health Research, Statistics, and Technology